Modélisation des Réarrangements Vα-Jα du TRA/TRD chez la souris et chez l'homme

Thuderoz Florence

UNIVERSITE JOSEPH FOURIER
SCIENCES. TECHNOLOGIE. MEDECINE. GRENOBLE - ALPES

TIMC-IMAG
TECHNIQUES DE L'INGÉNIERIE MÉDICALE ET DE LA COMPLEXITÉ

INFORMATIQUE, MATHÉMATIQUE ET APPLICATIONS GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE
Spécialité Igénierie de la santé, la cognition et l'environnement
Arrêté ministériel : 7 août 2006
Présentée et soutenue publiquement par

Thuderoz Florence

Le 4 novembre 2010

Modélisation des Réarrangements Vα-Jα du TRA/TRD chez la souris et chez l'homme

---

Thèse dirigée par **Jacques Demongeot**

JURY

Civilité/Nom/Prénom	Fonction et lieu de la fonction	Rôle
Christian Drouet	PU-PH, responsable de l'équipe GREPI, TIMC-IMAG, UMR CNRS 5525, La Tronche	**Président**
Charles Auffray	DR CNRS, Genexpress team – Systemoscope Consortium Functional Genomics and Systems Biology for Health	**Rapporteur**
Matthias Merkenshlager	Professor of Cell Biology, Institute of Clinical Science, Faculty of Medicine, Imperial College, Londres	**Rapporteur**
Marie-Paule Lefranc	PU, Directrice de l'équipe IMGT, IGH, UPR CNRS 1142, Montpellier	**Examinateur**
Patrice Marche	DR INSERM, Directeur de l'U823 INSERM, Institut Albert Bonniot, La Tronche	**Examinateur**

Thèse préparée au sein du Laboratoire TIMC (Techniques de l'imagerie, de la modélisation et de la cognition) dans l' Ecole Doctorale EDISCE (Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement)

Résumé

La recombinaison V(D)J constitue une recombinaison somatique et site spécifique de l'ADN à l'origine de la diversité des récepteurs antigéniques des lymphocytes T chez les vertébrés mandibulés. Concernant la chaîne a des récepteurs T, les gènes V et J sont utilisés depuis l'intérieur du locus TRA vers les gènes distaux durant des réarrangements successifs et ce sans exclusion allélique. La quantification expérimentale de certaines associations V-J chez la souris a permis de définir les tendances des répertoires combinatoires thymiques et périphériques. Un modèle numérique stochastique, basé sur des fenêtrages d'ouverture successives progressant sur les régions V et J durant les cycles de réarrangements, a permis une meilleure compréhension des règles dynamiques gouvernant les réarrangements V-J et a apporté la connaissance d'un répertoire combinatoire simulé renseignant les fréquences de toutes les associations V-J. Lors de la transition à l'homme, la quantification des associations V-J a été réalisée au niveau du thymus, constituant un premier échantillonnage à large échelle du répertoire combinatoire TRA humain. L'étape de modélisation a offert une compréhension claire de la construction dynamique du répertoire a humain et a permis de proposer des prédictions sur la diversité du répertoire combinatoire. Finalement, la progression de l'accessibilité des gènes aux réarrangements selon des vitesses d'ouverture non constantes associée à une ouverture synchronisée des régions J entre les deux allèles se sont révélées suffisantes pour expliquer les fréquences V-J expérimentales présentement disponibles pour les deux espèces ainsi que l'utilisation interallélique des gènes J.

T lymphocytes, in charge of cell mediated response, mostly express αβ antigen receptors at their surface. The T Cell Receptor (TCR) chain coding genes are formed through somatic site-specific recombinations, which associate Variable (V) and Junction (J) genes in developing lymphocytes for α chains (as well as Diversity (D) genes for β chains). The V(D)J recombinations allow jawed vertebrates forming an extremely large T Cell Receptor repertoire, beginning with the combinatorial association possibilities given by the several V, (D), and J genes available over chromosomal loci in germinal configurations. The physiologic combinatorial repertoire is well documented for β chains but not for α chains, due to a lack of available antibodies. More, the high number of Vα and Jα genes, the occurrence of successive rearrangements, and the absence of allelic exclusion increase the complexity of the involved mechanisms and limit the dynamical understanding of Vα-Jα rearrangements. Experimental quantifications of particular V-J associations were performed giving the tendencies of thymic and peripheral repertoires in mouse. Experiments along ontogeny allowed determining the empirical speeds of progression of V and J gene accessibility to rearrangements. A stochastic numerical model was designed determining successive opening windows that progress over the V and J regions during rearrangement rounds. The mouse model proposed in the thesis revealed new insights in the understanding of dynamical rules governing V-J rearrangements and provided a simulated combinatorial repertoire where appear frequencies for the entire V-J associations. In the transition to human, thymic quantifications of V-J associations were performed, providing a first experimental wide-ranging sampling of the human TCRα combinatorial repertoire. The modeling step, using a multi-level systemic approach followed by a simulation phase, offered a clear understanding of the dynamical building of the human α repertoire, in order to propose predictions on repertoire diversity richness and to dispose of a simulated repertoire showing the frequencies of the entire V-J associations. Eventually, both mouse and human models give a precise conservation of the mechanistic rules between the two species. Generation of peripheral experimental data is in progress, and preliminary results appear to validate the model approach for the human peripheral repertoire as well. Knowledge about repertoire shapes gives predictions on repertoire diversity richness and most of all constitutes an indispensable requirement in achieving the challenge of immune response characterization.

Présentation des réarrangements V(D)J

Les lymphocytes T sont des cellules immunitaires qui assurent des fonctions de vigilance à l'intérieur de l'organisme afin de détecter et d'éliminer les cellules cancéreuses ou infectées par un virus. Ces fonctions de reconnaissance reposent sur des récepteurs antigéniques exprimés à la surface du lymphocyte T. Les Récepteurs des Cellules T (TCR) sont des hétérodimères attachés à la membrane, majoritairement clonotypiques. Ces récepteurs reconnaissent spécifiquement des peptides étrangers ou du soi modifiés associés aux molécules du Complexe Majeur d'Histocompatibilité (CMH). Chaque chaîne des deux types d'hétérodimères αβ et γδ comprend un domaine constant et un domaine variable. La reconnaissance du complexe CMH-peptide est assurée par le domaine Variable qui contient trois régions hyper variables : les Régions Déterminantes de la Complémentarité (CDR1, 2 and 3). Les boucles CDR1 et CDR2 interagissent avec les molécules du CMH et CDR3 avec le peptide étranger. Les loci des chaînes de Récepteur T (TR) comprennent de multiples gènes de type Variable, (Diversité) et Jonction et ne sont pas fonctionnels à l'état germinal. La recombinaison V(D)J est un mécanisme sophistiqué de réarrangements somatiques sites spécifiques qui associe les segments de gènes durant des étapes spécifiques du développement lymphocytaire. Ce mécanisme produit des gènes fonctionnels codant les chaines de TR. Alors que CDR1 et CDR2 sont codés par deux régions distinctes du gène V, CDR3 correspond à la jonction V-J ou V-D-J. Cette jonction présente une diversité somatique très élevée à l'intérieur des populations lymphocytaires estimées à 10^8 et à 10^{12} lymphocytes chez la souris et l'homme respectivement. Le nombre de gènes V, (D) et (J) disponibles pour les réarrangements conduit à une diversité combinatoire. De plus, avant que les deux gènes qui réarrangent ne soient liés par la machinerie de réparation des extrémités non homologues de l'ADN, des nucléotides sont soustraits et ajoutés aléatoirement au niveau de la jonction codante, ce qui crée un diversité de jonction. L'association des deux chaînes de l'hétérodimère constitue un dernier facteur augmentant la diversité. L'immunocompétence étant supportée par la diversité du répertoire, la recombinaison V(D)J représente un mécanisme vital pour les vertébrés mandibulés.

T lymphocytes are immunity cells, which permanently check the entire organism to detect and eliminate virus infected or cancer cells. Recognition relies on antigenic receptors expressed at the T cell surface. T Cell Receptors (TCR) are membrane-bound heterodimers, mostly clonotypic, that specifically recognize foreign or modified peptides bounded to self-MHC (MHC for Major Histo-compatibility Complex). Each chain of the two heterodimer types, αβ and γδ, includes a constant and a variable domain. Peptide-MHC recognition is carried by the variable domain, which contains three highly variable Complementary-Determining Regions (CDR1, 2 and 3). CDR1 and CDR2 loops interact with self-MHC molecules and CDR3 with the foreign peptide. T Receptor (TR) loci encompass multiple different copies of Variable, (Diversity), and Junction genes and are non functional in germinal conformation. V(D)J recombination is a sophisticated somatic site-specific rearrangement mechanism associating gene segments during specific stages of lymphocyte development and generating functional genes for the T Cell Receptor chains. Although CDR1 and CDR2 are coded by two distinct regions over the V gene, CDR3 corresponds to the V-J or V-D-J junction. This junction displays an extensive somatic diversity in the 10^8 and 10^{12} lymphocyte populations estimated respectively in mouse and human. The number of V, (D), and J genes available to rearrangements provides combinatorial diversity. Additionally, before the two genes to rearrange are linked by Non Homologue End Joining process (NHEJ), random removal and non-templated addition of nucleotides at the coding joint creates N-region diversity or junction diversity. Pairing of heterodimer chains constitutes an ultimate factor increasing diversity. Immunocompetence relying on repertoire diversity, V(D)J recombination constitutes a vital mechanism for jawed vertebrates.

Presentation and interest of the modeling study

Jawed vertebrates T lymphocytes mostly express the αβ receptor on their surface. Diversity of corresponding β chains and the dynamical understanding of $V_\beta D_\beta J_\beta$ recombinations are well documented. Concerning α chains however, different complexities make difficult to address the dynamical aspects of the rearrangements and to understand the comprehensive combinatorial repertoire shape through an experimental approach by itself. The corresponding TRA (T Receptor Alpha) locus - made of the TRD locus, nestled in the TRA locus between the TRAV and TRAJ loci - is composed of numerous Vα and Jα genes, encompasses δ chain genes, and undergoes successive rearrangement rounds with no allelic exclusion. The successive windowing model developed is a stochastic numerical framework allowing the constitution of a TCRα combinational repertoire in a simulated lymphocyte population. This approach, after validation by comparing simulated with experimental results, allows making predictions on parameters values non accessible by experiments and definitely confirms that the progression of gene accessibility to rearrangements is a sufficient explanatory scenario for the TCRα combinatorial repertoire formation. Perspectives will be drawn concerning the role played by the genetic regulatory networks as well as concerning the influence of the chromatine dynamics in the control of the T cell differentiation pathways.

Ce mémoire se compose de 4 chapitres. Le premier chapitre resitue le mécanisme de réarrangement V(D)J dans la globalité du système immunitaire. Le second chapitre présente un état de l'art de la modélisation en immunologie, suivi d'une description d'un des premiers modèles. Dans ce second chapitre, sont présentées successivement une description générale de la modélisation des systèmes biologiques, suivie de ses applications en immunologie. Elles incluent la présentation de la modélisation des réarrangements Vα-Jα du locus TRA/TRD chez la souris, qui est ensuite discutée, dans une confrontation avec les données expérimentales. La transposition de ce modèle à l'homme figure au chapitre 3. La discussion et les perspectives développées à la vue des résultats constituent le quatrième chapitre. Une liste des abréviations et un glossaire des termes utilisés au long de ce mémoire se situent en page 13. La bibliographie est présente en fin de chacun des quatre chapitres.

Je vous souhaite une bonne lecture,

Plan du mémoire de thèse

Liste des abréviations / Glossaire

Antigène	substance déclenchant une réaction immunitaire
APC	cellule présentatrice de l'antigène (*Antigen-Presenting Cell*)
BCR	récepteur des cellules B (*B Cell Receptor*)
BEAD	élément de liaison alpha delta (*boundary element alpha delta*)
C	segment génique codant de type Constant présent sur les loci de récepteurs antigéniques
CDR	région déterminante de complémentarité
CLP	progéniteur commun aux lymphocytes (common lymphoid progenitors) ; le potentiel des CLP est restreint aux lymphocytes B, T et NK
CMH I	complexe majeur d'histocompatibilité de classe I
CMH II	complexe majeur d'histocompatibilité de classe II
CSB	bloc de séquences conservées (*Conserved sequence Block*)
CSH	cellules souches hématopoïétiques : petite population de cellules capables d'auto-renouvellement et multipotentes, à l'origine de la production de toutes les lignées cellulaires sanguines
D	segment génique codant de type Diversité, présent sur certains loci de récepteurs antigéniques
DN	Double Négatif
DNA-PK	Protéine kinase ADN-dépendante
DP	double positif
Eα	enhancer α
Eδ	enhancer δ
HLA	*Human Leucocyte Antigen*
HMG	groupe de haute mobilité (*High Hobility Group*)
HSC	cellules souches hématopoïetiques
HSP	protéines de choc thermique
Ig	immunoglobuline
IL	interleukine
ISP	cellule immature simple positif pour CD4+ ou CD8+
ITAM	motif d'activation du récepteur immunitaire, dépendant de la tyrosine (*immunoreceptor tyrosine-based activation motif*)
ITIM	motif d'inhibition du récepteur immunitaire, dépendant de la tyrosine (*immunoreceptor tyrosine-based inhibitory motif*)
J	segment génique codant, de type Jonction, présent sur les loci de récepteurs antigéniques
LCR	région de contrôle du locus (*locus control région*)
LPS	lipopolysaccharide
NK	cellules tueuses naturelles
RAG	enzyme activatrice de la recombinaison V(D)J
RSS	séquences signal de recombinaison
SP	simple positif
TCR	récepteur à l'antigène des cellules T
TdT	terminal désoxynucléotidyl transférase
TEA	promoteur précoce de la chaîne alpha (*T early* α)
TCF/LEF	facteur des cellules T (*T Cell Factor/lymphoid enhancer factor*)
TCR	récepteur des cellules T (*T Cell Receptor*)

TRAV	segment génique de type Variable de la chaîne alpha
TRBV	segment génique de type Variable de la chaîne beta
TRDV	segment génique de type Variable de la chaîne delta
TRGV	segment génique de type Variable de la chaîne gamma
V	segment génique codant de type Variable présent sur les loci de récepteurs antigéniques
XRCC4	X-ray cross complementation 4

Chapitre I Aspects biologiques du mécanisme de réarrangement V(D)J

Les systèmes immunitaires inné et acquis

L'immunité désigne l'ensemble des mécanismes de défense d'un organisme face à des éléments pathogènes (microorganismes : bactéries, parasites ou virus) ou des substances toxiques, ces éléments étant désignés sous le terme générique d'antigènes. Le système immunitaire assure également le maintien de l'intégrité de l'organisme contre des dégénérescences internes, telles que des mutations génétiques. Le système immunitaire des vertébrés mandibulés se compose d'une immunité innée et d'une immunité spécifique ou acquise.

Le système immunitaire inné ou immunité naturelle est une première ligne de défense commune à l'ensemble des organismes pluricellulaires. Dans le cas de ce système inné, la protection de l'organisme ne dépend pas du contact avec l'antigène, la réponse immunitaire est non spécifique et repose sur différents fronts de défense. L'organisme est en premier lieu défendu par les barrières externes physico-chimiques que sont la peau et les muqueuses. Si les antigènes viennent à pénétrer dans le corps, des facteurs chimiques solubles comme les enzymes bactéricides et la phagocytose (ingestion du pathogène par une cellule immunitaire) viendront compléter la réponse immunitaire innée. La reconnaissance des antigènes est immédiate et s'opère via des motifs communs à la plupart des microorganismes pathogènes : les *Pathogen Associated Molecular Patterns* (PAMP) ou *Microbe Associated Molecular Pattern* (MAMP). Le lipopolysaccharide bactérien (LPS), une endotoxine présente sur la membrane cellulaire bactérienne, est considéré comme le PAMP type. D'autres PAMP incluent la flagéline bactérienne, l'acide lipotéichoique des bactéries Gram positives, les peptides formylés (retrouvés dans l'ensemble du monde bactérien), les peptidoglycanes et les variants d'acides nucléiques normalement associés à des virus, comme l'ADN bactérien riche en séquences CpG déméthylées ou l'ARN double brin. La reconnaissance de ces motifs à l'intérieur de l'organisme déclenche très rapidement une réaction inflammatoire, qui vise à détruire et à exclure toute substance

étrangère de l'organisme. L'immunité naturelle est assurée par les cellules phagocytaires, les cellules dendritiques et le complément.

Le système immunitaire acquis, spécifique ou encore adaptatif est apparu il y a environ 500 millions d'années avec l'émergence des vertébrés mandibulés. Ses caractéristiques sont la spécificité vis à vis de l'antigène, la diversité, la mémoire et la reconnaissance du soi et du non soi. La réponse immune adaptative requiert un contact de l'organisme avec l'antigène pour être amorcée. Cette réponse est spécifique de l'antigène détecté et nécessite 3 à 5 jours pour être effective. Lorsqu'une réponse spécifique a été menée contre un antigène donné, la mémoire de cette réponse est conservée (cellules mémoire) permettant à l'organisme de mener une réponse plus efficace et plus rapide lors d'une deuxième rencontre. Le système immunitaire spécifique assure également le maintien de l'intégrité de l'organisme par l'élimination des cellules du soi modifiées. La réponse immune spécifique repose principalement sur deux types de cellules : les lymphocytes T et les lymphocytes B.

Les lymphocytes B et T

Les lymphocytes B et T se différencieraient à partir de progéniteurs cellulaires communs, les cellules souches lymphoïdes, elles-mêmes issues des cellules souches hématopoïétiques multipotentes ou hémocytoblastes de la moelle osseuses (présentes dans le foie durant la vie fœtale). Les lymphocytes qui demeurent dans la moelle osseuse et y poursuivent leur maturation se différencient en lymphocytes B, et ceux qui migrent vers le thymus (glande située dans le thorax, au-dessus de la crosse aortique cardiaque) se différencient en lymphocytes T (ou NK, *Natural Killer*). Les lymphocytes B et T rendus opérationnels vont migrer en périphérie pour assurer leur fonction de vigilance, qui repose sur la reconnaissance antigénique assurée par des récepteurs exprimés à leur surface. En cas d'activation, ils pourront s'associer à une réponse immune spécifique.

Les Récepteurs Antigéniques

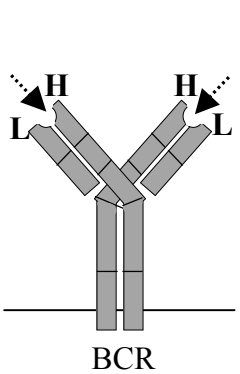
Les récepteurs antigéniques des lymphocytes B et T sont nommés respectivement BCR (*B Cell Receptor*) et TCR (*T Cell Receptor*). Un récepteur antigénique présente un ou plusieurs sites de reconnaissance à l'antigène, susceptible d'interagir de manière spécifique avec un motif antigénique ou épitope (Figures 1. A et C). Les récepteurs antigéniques sont majoritairement spécifiques d'un seul épitope et sont dits clonotypiques. Un lymphocyte donné produit ordinairement des récepteurs tous identiques et présente alors une spécificité antigénique unique. En revanche, la séquence, et donc la spécificité des récepteurs antigéniques, diffèrent d'un lymphocyte à l'autre. L'ensemble des lymphocytes de l'organisme présente alors des spécificités antigéniques très diverses, composant le répertoire immunitaire, et permet au système immunitaire de reconnaître tous les motifs antigéniques potentiellement existants. Lorsqu'un antigène est présent dans l'organisme, il va induire l'activation de lymphocytes spécifiques qui agiront pour l'éliminer. Une cellule du soi altérée déclenchera également l'activation de lymphocytes spécifiques qui contribueront à préserver l'intégrité de l'organisme.

Les BCR ou immunoglobulines (Ig) sont formées de deux chaînes protéiques lourdes et de deux chaînes légères (Figure 1.A). Il existe deux types de chaînes légères, Igκ et Igλ, ainsi qu'un type de chaîne lourde, IgH. Les loci codant les chaînes IgH, Igκ et Igλ se situent respectivement

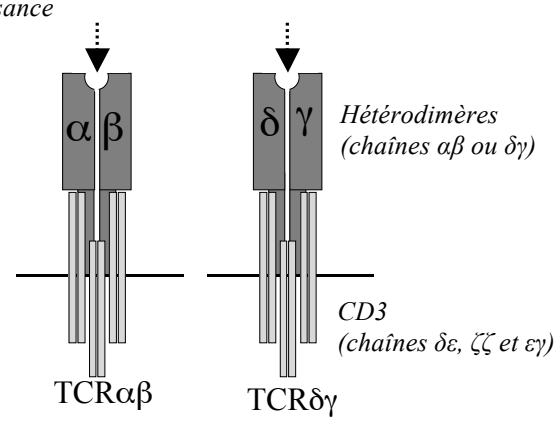
sur les chromosomes 14, 2 et 22 chez l'Homme, et sur les chromosomes 12, 6 et 16 chez la Souris (Figure 1.B).

Les TCR sont des hétérodimères de type TCRαβ ou TCRδγ (Figure 1.C). Les chaînes protéiques les composant sont synthétisées à partir des loci TRA/TRD, TCRB et TCRG, situés respectivement sur les chromosomes 14, 7 et 7 chez l'homme et sur les chromosomes 14, 6 et 13 chez la souris (Figure 1.D). Alors que les lymphocytes Tαβ sont majoritaires dans le sang, les cellules Tδγ sont très abondantes dans les muqueuses de l'organisme. Chaque chaîne de TCR comporte une région transmembranaire contenant des résidus aminés hydrophobes, ainsi qu'une courte queue cytoplasmique ne permettant pas la transduction d'un signal. Le TCR est associé à un groupe de 6 protéines (chaînes ε, δ, ξ, ξ, γ et ε) collectivement nommées CD3 à la surface du lymphocyte T. Ce complexe multiprotéique, nécessaire à l'expression du TCR à la surface cellulaire durant le développement lymphocytaire, a pour fonction de transduire le signal d'activation au lymphocyte, lorsque le TCR est stimulé.

B. Schéma du BCR, récepteur antigénique du lymphocyte B



C. Schéma du TCR, récepteur antigénique du lymphocyte T



A. Localisation chromosomique des différents loci porteurs des gènes codant pour les chaînes de BCR

Chaînes de BCR	Homme	Souris
gènes des chaînes lourdes Locus IGH	Chromosome 14 (14q32.33)	Chromosome 12 (F2 58.0-59.0cM)
gènes des chaînes légères κ Locus IGK	Chromosome 2 (2p11.2)	Chromosome 6 (C2 30.0cM)
gènes des chaînes légères λ Locus IGL	Chromosome 22 (22q11.2)	Chromosome 16 (B1 13cM)

D. Localisation chromosomique des différents loci porteurs des gènes codant pour les chaînes de TCR

Chaînes de TCR	Homme	Souris
Gènes des chaînes TCRα et TCRδ Locus TRA/TRD	Chromosome 14 14q11.2	Chromosome 14 19.7cM
Gènes des chaînes TCRβ Locus TRB	Chromosome 7 7q34	Chromosome 6 6A-C
Gènes des chaînes TCRγ Locus TRG	Chromosome 7 7p14	Chromosome 13 13A2-3

Figure 1. A : schéma d'un récepteur BCR à la surface d'un lymphocyte B. **B** : localisation chromosomique des loci codant pour les différentes chaînes de BCR. **C** : schéma d'un récepteur TCR à la surface d'un lymphocyte T. Le TCR est composé d'un hétérodimère et d'un complexe protéique CD3. **D** : localisation chromosomique des loci codant pour les différentes chaînes de TCR. Source pour la localisation des loci : www.imgt.fr

CDR et diversité de reconnaissance

Les chaînes lourdes de BCR et les chaînes de TCR comportent chacune une région invariante dans sa séquence et proche de la membrane, dite domaine constant. De plus, toutes les chaînes de récepteurs antigéniques présentent un domaine hautement variable, distal par rapport à la membrane, appelé région variable. La variabilité de ce dernier domaine se concentre sur trois boucles situées à l'extrémité de la chaîne la plus éloignée de la membrane, les régions hypervariables ou régions déterminantes de la complémentarité (CDR, *Complementarity Determining Regions*). Les CDR sont relativement courts en séquences d'acides aminés (Figure 2) et déterminent la spécificité du BCR ou du TCR pour son ligand. Parmi les trois CDR, CDR3 présente la plus extrême variabilité. La capacité de l'organisme à reconnaître tous les antigènes potentiellement existants repose sur la diversité des domaines variables des chaînes de récepteurs antigéniques et plus précisément sur la variabilité de séquence des CDR.

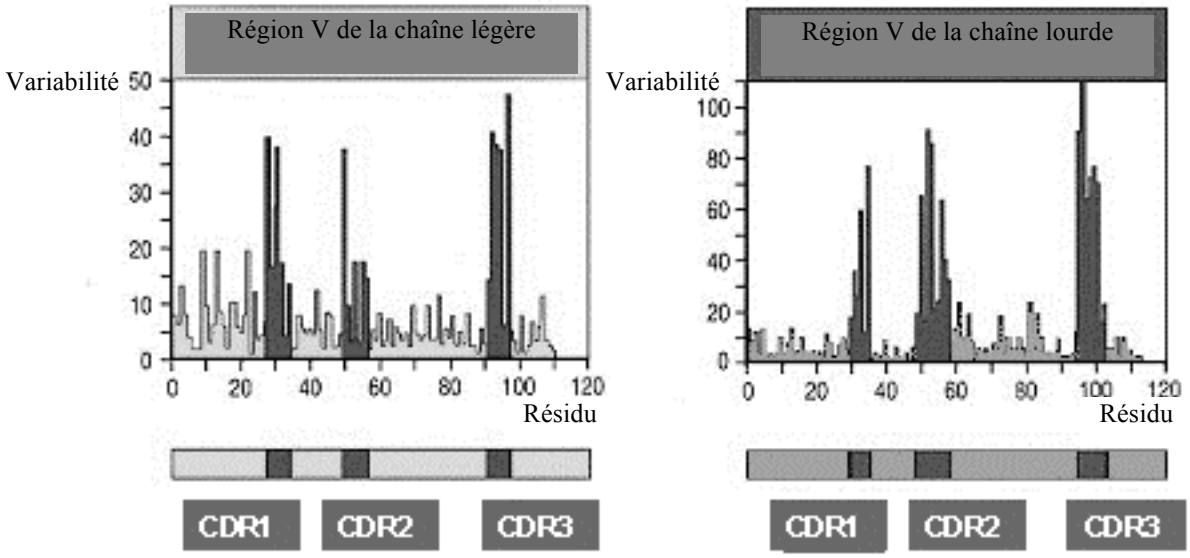


Figure 2. Régions CDR d'une chaîne légère et d'une chaîne lourde d'immunoglobuline. La variabilité est mesurée en divisant, pour chaque position, le nombre d'acides aminés différents répertoriés par le nombre de fois où le résidu le plus fréquent a été observé.

Activation des Lymphocytes B

Suite à la pénétration d'un antigène dans l'organisme ou à l'altération de cellules du soi, un certain nombre de lymphocytes spécifiques vont être activés. Ces lymphocytes activés vont se différencier et se multiplier pour répondre à l'infection, dans le but de rétablir l'intégrité de l'organisme. Les modes d'activation diffèrent en fonction du type lymphocytaire considéré. Les lymphocytes B répondent à une infection d'antigènes libres, alors que les lymphocytes T interviennent lorsqu'une cellule de l'organisme est altérée. Les lymphocytes Tαβ et Tδγ ne sont

pas activés selon le même mode et, parmi les lymphocytes Tαβ, sont encore distingués les lymphocytes $CD4^+$ et $CD8^+$, en raison de la présence de molécules marqueuses de différenciation à leur surface (*Cluster Differentiation*).

L'action des lymphocytes B est dirigée contre les antigènes dits libres, c'est-à-dire n'ayant pas pénétré une cellule de l'organisme. Les lymphocytes B sont activés par contact direct avec leur antigène spécifique. Une fois activés, ils se divisent fortement (expansion clonale), puis se différencient en plasmocytes. Ils produisent et sécrètent alors la forme libre du BCR : l'anticorps. Les anticorps se fixent directement sur l'antigène ; l'élément pathogène se trouve alors littéralement recouvert d'anticorps neutralisants, qui vont faciliter sa phagocytose par le mécanisme d'opsonisation. Ce recouvrement va également activer l'élimination du pathogène par le complément. Cette immunité portée par les molécules d'anticorps qui transitent par le sang est désignée par le terme d'immunité à médiation humorale. L'immunité à médiation humorale permet de répondre à une infection en phase précoce : cette réponse a donc une cible extracellulaire et agit avant que le pathogène ne pénètre les cellules de l'organisme.

Activation des lymphocytes Tαβ

Les lymphocytes Tαβ de type $CD4^+$ ou $CD8^+$ ne reconnaissent pas directement l'antigène. Ils sont activés par les cellules présentatrice d'antigène du système immunitaire (APC, *Antigen-Presenting Cells*). Ces cellules présentatrices vont ingérer l'antigène, lyser ses constituants (apprêtement de l'antigène ou *processing*) et vont arborer à leur surface un peptide antigénique [Shimonkevitz *et al*, 1983]. Ce peptide est porté par une molécule du Complexe Majeur d'Histocompatibilité (CMH) nommée également molécule du système HLA (*Human Leucocyte Antigen*) (Figure 2.B). Un TCR reconnaît l'association molécule du CMH/peptide antigénique. Les lymphocytes T dont le TCR est spécifique de l'une des associations CMH/peptide exogène présentées seront activés. L'interaction entre le TCR et la molécule du CMH étant faible, l'engagement d'autres molécules est nécessaire à sa stabilisation, l'ensemble formant la synapse immunologique.

Pour permettre l'activation d'un lymphocyte Tαβ, les molécules du CMH doivent également interagir avec les marqueurs de différenciation CD4 ou CD8, qui constituent des molécules de co-activation. Il existe deux types de molécules de CMH présentes à la surface des cellules présentatrices d'antigènes : les molécules du CMH-I et du CMH-II. La molécule de co-activation CD8 reconnaît les molécules du CMH-I, tandis que CD4 interagit avec les molécules du CMH-II (Figure 3). Les lymphocytes T $CD4^+$ ne pourront être activés que lorsque le peptide antigénique est présenté en association avec les molécules du CMH-II et les lymphocytes T

$CD8^+$ seront activés au contraire lors de la présentation de peptides exogènes portés par les molécules du CMH- I.

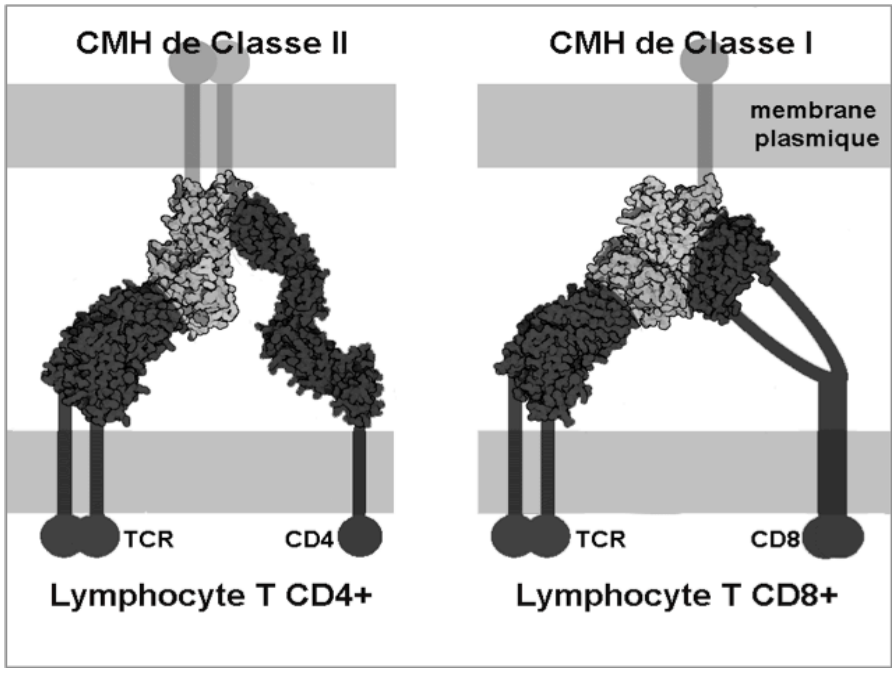


Figure 2. Reconnaissance du complexe CMH/peptide antigénique par le TCR/co-récepteur. Reconnaissance du CMH-I par la molécule de co-activation CD8 et du CMH-II par CD4. Source : Molecule of the Month by www.pdb.org

Dans l'organisme, Les molécules du CMH-I sont présentes sur toutes les cellules nucléées. Lorsqu'une cellule est saine, elle arbore sur ses molécules du CMH-I un peptide de l'organisme dit du soi ; quand elle est infectée ou modifiée de manière pathologique (cellule cancéreuse), elle va présenter un peptide antigénique ou un peptide du soi-modifié. Les lymphocytes T $CD8^+$ reconnaissent les associations CMH-I/peptide antigénique ou CMH-I/peptide du soi-modifié (Table 1). Toutes les cellules de l'organisme sont ainsi en permanence soumises à un contrôle par les lymphocytes T. Lorsqu'ils sont activés, les lymphocytes Tαβ $CD8^+$ de nature cytotoxique (CTL, *Cytotoxic T Lymphocytes*) vont se diviser activement et entraîner la lyse de leurs cellules cibles reconnues comme anormales, permettant à l'organisme de retrouver son intégrité.

Les molécules du CMH-II ne sont présentes que sur les cellules immunitaires. Plus particulièrement, les monocytes, les macrophages, les cellules dendritiques et les lymphocytes B expriment de manière constitutive les molécules de classe II du CMH (Table 1). D'autres types cellulaires peuvent également assurer une présentation du peptide antigénique en association avec les molécules de classe II du CMH, comme les cellules épithéliales de l'intestin, les kératinocytes, les hépatocytes, les astrocytes et les cellules de la microglie du système nerveux central. Les lymphocytes T $CD4^+$ reconnaissent les associations CMH-II/peptide antigénique ou CMH-II/peptide du soi-modifié (Table 1). L'activation de ces lymphocytes T dits auxiliaires (LTh, *T-helper*) est un événement crucial dans l'induction d'une réponse immune : ces cellules vont proliférer et activer en quantité d'autres types cellulaires, qui agiront de manière plus directe dans la réponse. Cette régulation est effectuée via la sécrétion d'interleukines, molécules messagères spécifiques du système immunitaire.

Molécule du CMH	Type de cellule portant ces molécules du CMH	Reconnaissance par les molécules de co-activation
CMH-I	Toutes les cellules nucléées de l'organisme	CD8
CMH-II	Cellules immunitaires	CD4

Table 1. Types cellulaires portant les molécules du CMH-I et CMH-II et reconnaissance par les molécules de co-activation CD8 et CD4.

Finalement, les molécules accessoires CD2 et LFA-1, présentes en surface des lymphocytes T αβ, interagissent avec LFA-3 et ICAM-1, portées par la cellule présentatrice d'antigène, et viennent également renforcer l'interaction TCR/CMH (LFA : *Leukocyte Function-associated Antigen*, ICAM: *Intercellular Adhesion Molecule*). La Figure 4 résume la composition du complexe synaptique de reconnaissance pour un lymphocyte T αβ $CD4^+$.

En plus des molécules accessoires qui stabilisent l'interaction entre le TCR et l'antigène associé à la molécule du CMH, la synapse immunologique fait intervenir des molécules co-activatrices, comme CD28, qui doivent être engagées avec leurs ligands pour permettre l'activation du lymphocyte T. Les molécules accessoires et co-activatrices sont invariantes et ne contribuent en aucun cas à la spécificité de la reconnaissance.

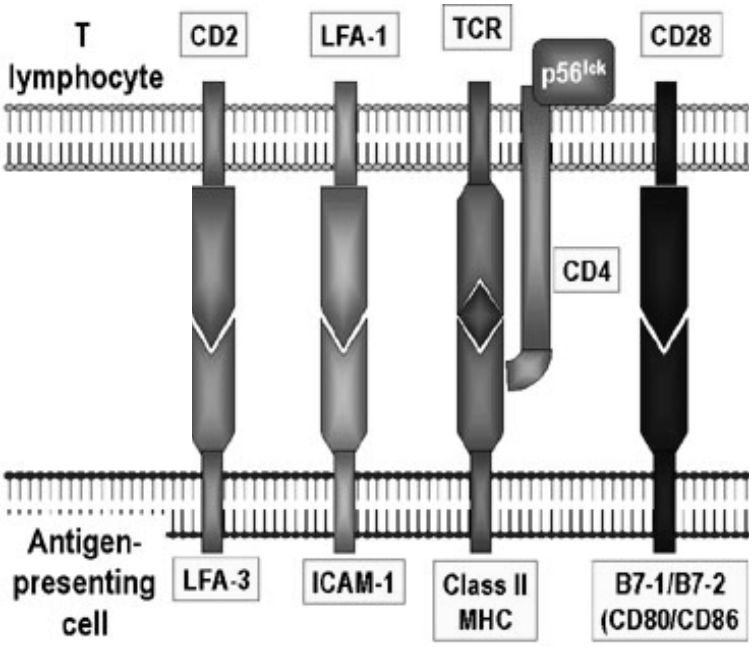


Figure 4. Schéma illustrant la composition de la synapse immunologique d'un lymphocyte T αβ CD4+. Image provenant du site de l'Université de Médecine de Caroline du Sud (pathmicro.med.sc.edu)

Activation des lymphocytes T δγ

Le pourcentage des lymphocytes Tδγ est variable selon les espèces. Chez les moutons, bovins, porcs et poulets, ce type cellulaire représente la moitié des lymphocytes sanguins (*"γδ high species"*) [Hein *et al*, 1991; Cooper *et al*, 1991]. Chez l'homme et la souris, les lymphocytes Tδγ sont moins abondant que les Tαβ et représentent environ 5% du total des Lymphocytes T sanguins (*"γδ low species"*), néanmoins, ce type lymphocytaire prédomine dans les muqueuses épithéliales, parmi la population des lymphocytes intra-épithéliaux [Hayday *et al*, 2000]. Les Lymphocytes Tδγ ne portent pas, pour la plupart, de marqueurs CD4 ou CD8, et leur activation est indépendante des molécules du CMH. Leurs TCR ne semblent pas être restreints à la reconnaissance d'un peptide, mais pourraient réagir à la présence d'une protéine entière [Herzig *et al*, 2006]. Les lymphocytes Tδγ présentent un répertoire biaisé, dirigé contre certaines bactéries et antigènes viraux, et leur activité est généralement cytotoxique.

Pour conclure, la fonction d'un lymphocyte repose sur sa faculté à reconnaître spécifiquement un antigène. Cette reconnaissance est permise par les récepteurs antigéniques (BCR et TCR). Le mécanisme de recombinaison V(D)J, sujet d'étude de cette thèse, permet de

générer des gènes fonctionnels d'une grande diversité, pour la synthèse des différentes chaînes de récepteurs antigéniques. Les réarrangements des loci codant pour les différentes chaînes de TCR ont lieu à des étapes spécifiques du développement du lymphocyte T.

La lymphopoïèse T

La lymphopoïèse T fait intervenir une migration. La production et l'amplification de progéniteurs immatures, à partir des cellules souches hématopoïétiques, est assurée par le foie durant la vie fœtale, puis par la moelle osseuse. Ces progéniteurs colonisent le thymus pour y poursuivre leur différentiation, qui inclue les réarrangements de loci codant pour les chaînes de TCR, un processus de sélection positive et négative, et une étape finale de maturation. Les lymphocytes T opérationnels ainsi produits migreront à la périphérie pour regagner la population de lymphocytes naïfs, lymphocytes matures n'ayant pas encore rencontré l'antigène. Ces lymphocytes assureront en périphérie leur fonction de vigilance et, s'ils sont activés, mèneront leurs fonctions effectrices dans la réponse immune spécifique à médiation cellulaire.

Les cellules souches hématopoïétiques constituent une petite population de cellules qui représente 0,01 à 0,05% des cellules médullaires. Ces cellules souches hématopoïétiques (CSH) sont capables d'auto-renouvellement et sont pluripotentes : elles se placent à l'origine de la production de toutes les cellules sanguines [Till et Mc Culloch, 1961]. Les CSH se différencient en passant par une série de progéniteurs intermédiaires. Plus une cellule hématopoïétique s'engage sur la voie de la différenciation, plus sa capacité d'auto-renouvellement et sa multi-potentialité diminuent. Dans la hiérarchie hématopoïétique, les progéniteurs intermédiaires constituent des points de branchement et de bifurcation entre les différentes lignées cellulaires sanguines. La décision de la destinée d'une cellule s'effectue à travers un réseau d'expression de gènes et de changements épigénétiques. Les CSH sont à l'origine des lignées myéloïdes (érythroïde, mégacaryocytaire et granulomacrophagique) et lymphoïdes (LB, LT et NK). Cependant, l'identité du progéniteur intermédiaire qui constitue le point de branchement entre les lignées lymphocytaires T et B ne fait pas consensus. Classiquement, le progéniteur lymphoïde commun (CLP) est considéré comme le point de branchement, à partir duquel toutes les lignées lymphoïdes émanent [Kondo et al. 1997]. Des recherches plus récentes continuent d'appuyer cette hypothèse [Akashi, 2000 ; Karsunky et al. 2008]. Néanmoins, selon d'autres auteurs, les CLP se différencieraient majoritairement en lymphocytes B et les progéniteurs des lymphocytes T issus de la moelle osseuse correspondraient majoritairement à une population située en amont des CLP, dans la hiérarchie hématopoïétique, constituée des progéniteurs multipotents (MPP) [Schwarz and Bhandoola 2004; Perry et al. 2006; Lai and Kondo 2007; Wada et al. 2008]. Si la nature des progéniteurs hématopoïétiques migrant du foie fœtal ou de la moelle osseuse vers le

thymus n'est pas encore clairement identifiée, néanmoins, à l'entrée dans le thymus, les cellules potentielles lymphoïdes B et myéloïdes résiduels sont rapidement perdues et les cellules potentielles NK, cellules dentritiques et macrophages, persistent temporairement [Balciunaite et al., 2005 ; Bhandoola et al., 2007]. Ainsi, durant la spécification T, les cellules conservent une variété d'options de différentiation (mais également une forte capacité de prolifération), au cours de la mise en place de leur engagement dans le lignage T et ce, même après le choix de leur destin cellulaire. Les progéniteurs T vont se différencier selon des étapes précises tout au long de leurs déplacements à travers les différents compartiments anatomiques du thymus. Les cellules progénitrices thymiques immatures passent ainsi successivement par les stades Double Négatif (DN), Immature Simple Positif (ISP), Double Positif (DP) et Simple Positif (SP) (Figure 5).

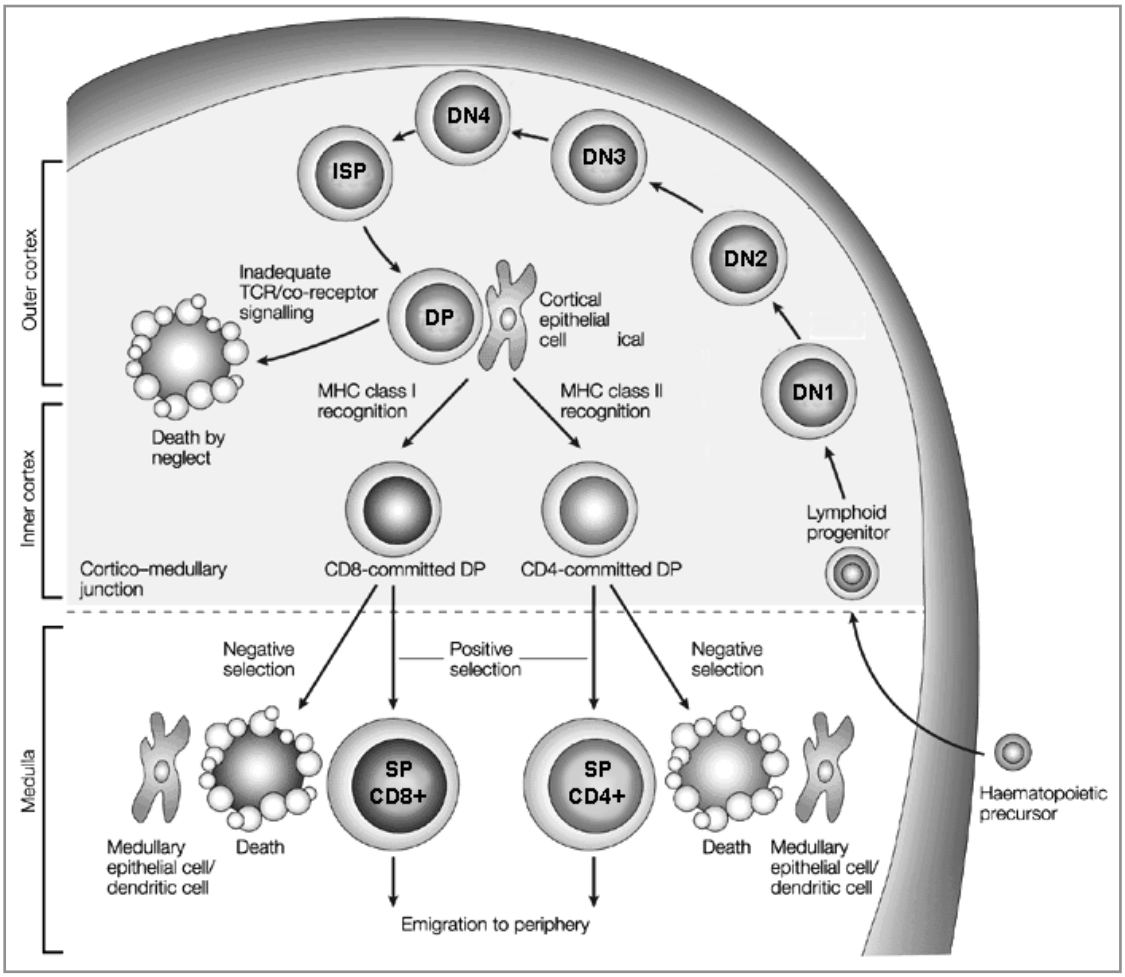


Figure 5. Schéma illustrant les stades de développement des lymphocytes T en transition dans le cortex externe, le cortex interne, puis dans la medulla thymique. Les stades Double Négatif (DN), Immature Simple Positif (ISP), Double Positif (DP) et Simple Positif (SP) sont présentés (D'après [Ronald N. Germain T-cell development and the CD4–CD8 lineage decision Nature Reviews Immunology 2, 309-322 (May 2002)])

Les Cellules Double Négatives

Les cellules Double Négative (DN) présentent toutes un phénotype CD4-CD8-. Chez la souris, 4 fractions sont encore distinguées (selon l'expression des antigènes CD25 et CD44) : alors que les cellules DN1 et DN2 sont nommées progéniteurs (Pro-T), les stades DN3 et DN4 sont des précurseurs (Pré-T). Le terme progéniteur signifie cellule souche différenciée ou engagée dans la différenciation ; en avançant dans leurs étapes de spécification, ces cellules en viennent à s'engager dans une lignée cellulaire unique et sont alors désignées par le terme de précurseurs.

Les sous-populations de cellules immuno-compétentes sont identifiées phénotypiquement sur la base de la combinaison des antigènes qu'elles expriment à leur surface. Ces antigènes, remplissent des fonctions de signalisation ou d'adhésion pour la cellule. Ils sont utilisés comme marqueurs, + : présence de l'antigène
- : absence de l'antigène
Exemples :
CSH : CD34+, CD31-
Tous les groupes de leukocytes : CD45+
Lymphocytes T : CD45+, CD3+
Lymphocytes B : CD45+, CD19+ ou D45+, CD20+

Les précurseurs intra-thymiques les plus primitifs et possédant la plus grande capacité de prolifération présentent un phénotype c-Kit+ DN1 (Figure 6) [Ceredig and Rolink, 2002; Porritt et al., 2004]. Les cellules DN1 sont CD117+CD44+CD25- : ces progéniteurs très immatures possèdent encore un potentiel lymphoïde multiple (B et NK) et myéloïde résiduel [Godfrey et al, 1993]. A ce stade, aucun réarrangement de loci codant pour les chaînes de TCR n'est effectué : ces progéniteurs présentent encore des expressions de gènes du lignage T très équivoques.

Le stade DN2 marque la perte du potentiel lymphoïde et du potentiel myéloïde résiduel. Les cellules DN2 produiront majoritairement des lymphocytes T αβ et T γδ (également des NKT et Treg). Ces progéniteurs sont c-Kit+CD117^{High}CD44+CD25+ : ils présentent une augmentation de l'expression de multiples gènes du lignage T, même si le complexe Rag1/2, responsable de l'initiation des réarrangements des gènes de chaînes de TCR, est encore faiblement exprimé. Le réarrangement des loci codant pour les chaînes de récepteurs antigéniques TCRδ et TCRγ s'initient durant cette phase. Les stades DN1 et DN2 se caractérisent par une prolifération intense, qui chute au stade DN3.

Les cellules DN3, de phénotype c-Kit-CD117^{low}CD44-CD25+, présentent une très faible prolifération et un fort taux de réarrangement. Ces cellules réarrangent fortement le locus codant pour la chaîne β, tout en continuant de réarranger les loci des chaînes TRδ et TRγ [Tourigny et al. 1997; Capone et al. 1998]. Si les cellules DN3 continuent d'exprimer quelques gènes distinctifs qui marquent leur caractère immature, elles expriment sans équivoque tous les gènes nécessaires à la signalisation T. Dans ces cellules, un changement intrinsèque déterminant se

produit au niveau du réseau de régulation génique qui stoppe l'accès aux lignages alternatifs Natural Killer, cellules dendritiques et macrophages, quelles que soient les conditions de signalisation appliquées. Un point de contrôle se situe également à ce stade de développement des lymphocytes T : les cellules ne recouvreront une activité de prolifération, que si elles complètent avec succès leurs réarrangements. Ce stade marque aussi la division entre les lignées Tαβ et Tγδ. Si le réarrangement du locus de la chaîne β est productif, cette chaîne protéique sera synthétisée, et s'associera avec la chaîne pTα (chaîne α immature) et le complexe CD3, pour former le récepteur antigénique immature, le pré-TRC. La transition des stades DN3 et DN4 marque l'expression en surface du récepteur pré-TRC : cette transition correspond, chez la souris, à l'étape d'élimination par apoptose des cellules n'ayant pas réarrangé correctement leurs locus de chaîne β ; c'est la sélection β [Wilson, Marechal, 2001].

Le stade DN4 correspond à des cellules CD117^{low}CD44-CD25-. A ce stade, le pré -TR est significativement exprimé à la surface cellulaire, et peut être détecté [Godfrey et al. 1993]. Les cellules DN4 prolifèrent fortement.

Chez l'homme, c'est l'expression du marqueur CDα qui signe le passage d'un progéniteur encore pluripotent à un stade clairement engagé dans la lignée lymphocytaire T.. La sélection β pourrait correspondre à la transiton DN-ISP [Dik et al. 2005].

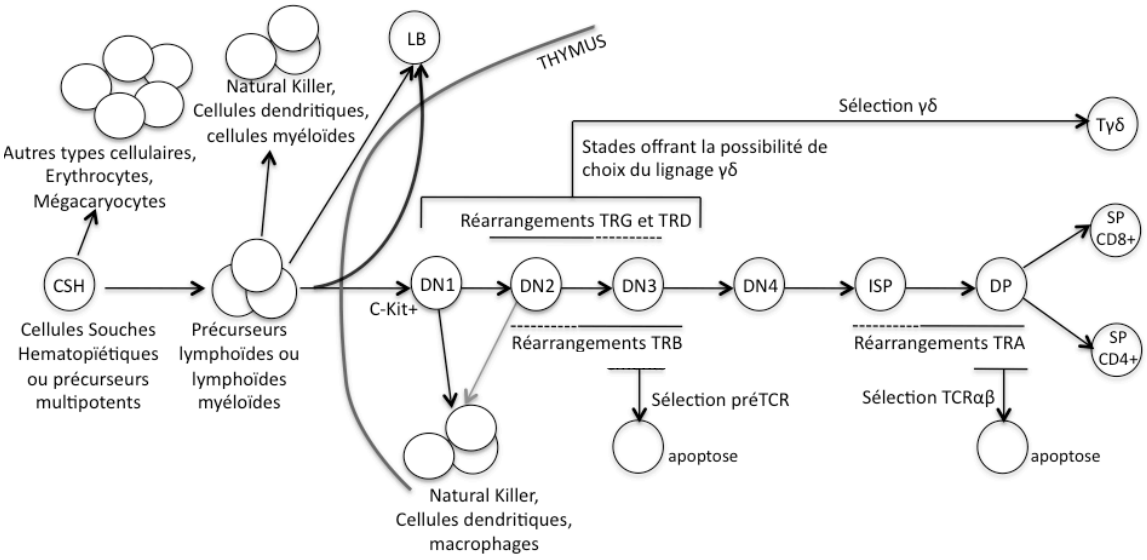


Figure 6. Etapes du développement des lymphocytes T chez la souris. Les réarrangements des différents loci des chaînes de TCR sont indiqués par des lignes pleines, lorsque ces réarrangements sont majoritaires, et en pointillés quand les réarrangements sont faibles. Les possibilités de lignage persistant à chaque stade de développement sont indiquées.

cellules Immatures Simple Positive

Les cellules Immatures Simple Positive (ISP) expriment faiblement le co-récepteur CD4int ou CD8int. Ce stade marque l'initiation des réarrangements du locus codant pour la chaîne TCRα.

Les cellules Double Positive

Les cellules Double Positive (DP) présentent un phénotype CD4+CD8+. Il s'agit du stade durant lequel les gènes codant pour la chaîne TCRα vont majoritairement se réarranger. Les cellules DP vont être progressivement capables d'exprimer le TCRαβ à leur surface [Pearse, et al. 1989], en passant par les stades TCR^{low}, TCR^{int} et TCR^{High}. Les cellules TCR^{low} proviennent directement des cellules ISP. Les cellules TCR^{int} présentent un niveau d'expression du TCR de densité intermédiaire : ce stade correspondrait à des cellules venant de terminer l'appariement de leurs chaînes αβ et soumettant leurs TCRαβ, nouvellement exprimés en surface, à la sélection thymique ((le processus de sélection thymique sera détaillé dans un prochain paragraphe).). Les cellules TCR^{High}, présentant une forte densité de TCRαβ en surface, correspondraient aux cellules ayant passé la sélection thymique avec succès.

Les cellules Simple Positive

Les cellules Tαβ Simple Positive (SP) sont soit CD4+, soit CD8+. L'engagement d'un lymphocyte T dans la lignée SP CD4+ ou SP CD8+ serait déterminé en fonction de la reconnaissance du CMH-I ou du CMH-II par le TCR, lors de la sélection thymique positive [Kaye et al. 1989]. Les cellules SP se caractérisent par une prolifération qui redevient élevée et effectuent en moyenne 6 cycles de divisions [Ceredig et al, 1990; Hare et al. 1998].

Les lignages Tαβ et Tγδ

Le locus TRA/TRD comprend les gènes codant pour les chaînes α et δ. Au sein de ce locus, les gènes codant pour la chaîne δ sont disposés entre les gènes de la chaîne α, si bien qu'un réarrangement du locus TRA subtilise l'ensemble du locus TRD. La structure composée des loci TRA/TRD impose donc que les recombinaisons TRD soient réalisées précédemment aux réarrangements du locus TRA. Expérimentalement, les réarrangements des gènes codant pour les chaînes TCRδ et TCRγ sont observés de manière très précoce, dès le stade DN2 et les lymphocytes T γδ sont détectés à partir du treizième jour fœtal [Capone *et al.* 1998]. Les loci

codant pour chaîne TCRα ne commencent à être réarrangés qu'à partir du stade DN4, les cellules Tαβ n'étant détectables qu'à partir du 17ème jour de gestation [Mancini et al. 1999]. Le choix cellulaire du lignage Tγδ serait déterminé précédemment aux réarrangements des différents loci : en cas de réarrangement non productif, la cellule aurait la possibilité de s'engager vers le lignage αβ.

Choix du lignage lymphocytaires T

De manière générale, la différentiation peut s'opérer par l'activation de gènes spécifiques d'un type cellulaire. Dans ce cas, la cellule qui suit un chemin de différentiation est exclue de toute autre alternative d'expression génique et donc de tout autre choix de type cellulaire par 3 principaux mécanismes : l'autorégulation positive des facteurs principaux de transcription spécifiques du type cellulaire, l'interaction de ces facteurs principaux avec d'autres complexes moléculaires, et l'action antagoniste exercée contre les facteurs spécifiques d'autres lignées cellulaires. Ces mécanismes enclenchent des réactions en cascade dans les réseaux de régulation, qui deviennent irréversibles et instaurent une frontière marquant définitivement le choix de lignage de cette cellule. Ainsi, la différentiation des lymphocytes B repose sur l'expression de facteurs de transcription spécifiques du lignage (PU.1 et Ikaros), qui déclenchent des réactions en cascades (Figure 7) [Singh et al. 2005].

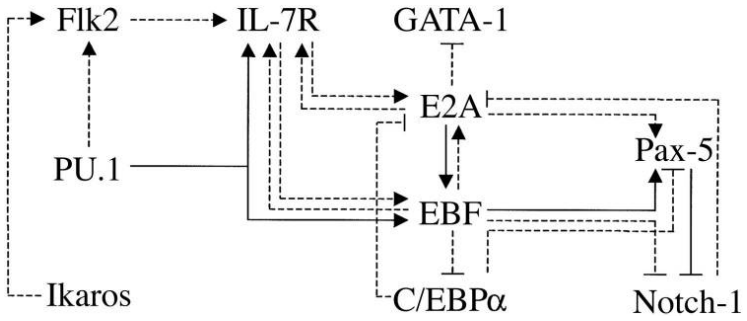


Figure 7. Réseau d'interactions auto-entretenu, établi pour un précurseur lymphocytaire B. Les effets activateurs et inhibiteurs sont représentés en lignes pleines ou pointillées, selon le niveau d'évidence expérimentale. L'établissement de cet état du réseau d'interactions dépendrait de l'activation transitoire du récepteur aux cytokines Flk2 et des facteurs de transcription PU.1 et Ikaros. Des boucles de rétroaction positive impliquent le récepteur aux cytokines IL-7R et les facteurs de transcription EBF, E2A et Pax-5 pourraient générer le circuit auto-entretenu. L'architecture du réseau suggère un antagonisme croisé de facteurs de transcription déterminant le destin cellulaire (GATA-1, C/EBPα, and Notch-1). Figure et légendes extraites de [Singh, 2005].

La spécification d'un type cellulaire peut aussi reposer sur l'équilibre dynamique de

facteurs de régulation non spécifiques du type cellulaire. Ainsi, la différentiation T semble ne pas impliquer de facteurs spécifiques. L'identité lymphocytaire T se forme principalement à travers l'implication de 9 facteurs de transcription, majoritairement non spécifiques du lignage (GATA-3, famille de facteurs TCF, les ''E protéines'' E2A ou HEB, famille de facteurs Ikaros, Myb, Gfi-1, famille de facteurs Runx, famille de facteurs Ets PU.1, et CSL) [Rothenberg and Taghon, 2005 ; Blom and Spits, 2006]. Ces facteurs se placent sous l'influence du chemin de signalisation de Notch-Delta provenant du micro-environnement thymique (activation par les ligands DL1 ou DL4 exprimés par les cellules du stroma thymique) [Besseyrias et al., 2007]. Le facteur Notch-Delta occupe une place spéciale, parmi les autres facteurs précédemment listés, car il constitue le seul élément régulateur capable de diriger une cellule hématopoïétique encore non engagée vers le programme de développement lymphocytaire T. Après cet engagement, le facteur Notch-Delta agit également de manière répétée, pour maintenir les cellules dans les stades de développement pro-T [Maillard et al., 2005]. Cette forte influence sur le lignage T est particulière, car Notch-Delta joue par ailleurs des rôles divers dans le développement embryonnaire [Bray, 2006; Artavanis-Tsakonas et al., 1999]. Dans le choix du lignage T, Notch-Delta agit au sein d'un réseau de régulation complexe, et à travers différents mécanismes, ce facteur devant interagir avec un ensemble de cibles variant à chaque stade du développement T [Georgescu, 2008]. En bref, les cellules c-Kit+ DN1 présentent rapidement une nécessité de la signalisation Notch pour leur génération et leur maintien, cette dépendance se perpétuant jusqu'au stade des thymocytes DN3, qui nécessitent ce signal pour leur survie, mais également pour leur compétence à suivre la sélection Beta. L'émancipation de Notch intervient à la suite des sélections β ou γδ, lorsque l'engagement cellulaire est sur le point d'être totalement complet [Georgescu, 2008].

La recombinaison V(D)J

La recombinaison V(D)J est un mécanisme de réarrangement somatique et site-spécifique de l'ADN, qui est à l'origine de l'extraordinaire variabilité structurale des sites de reconnaissance à l'antigène des immunoglobulines et des TCR.

Les Gènes Variable, Diversité, Jonction et Constant

Les loci codant pour les chaînes de récepteurs antigéniques des lymphocytes sont non fonctionnels au stade germinal ; ils ont en commun de présenter des segments codant multiples, séparés par des séquences introniques. Ces séquences codantes sont nommées gènes de type Variable (V), Diversité (D), Jonction (J) et Constant (C). Les loci des chaînes de TCR (TRA, TRB, TRG et TRD) et des chaînes de BCR (IGH, IGK et IGL) comportent chacun un nombre différent de ces gènes, les segments de types Diversité n'étant pas présents sur tous les loci.

Durant la différenciation lymphocytaire, la recombinaison V(D)J des loci spécifiques permet d'assembler un segment codant de chaque type V, (D) et J, afin d'obtenir un gène fonctionnel pour la synthèse d'une chaîne de récepteur antigénique [Bassing *et al*, 2002]. Les segments V(D)J, alors rassemblés en un seul exon, vont pouvoir être transcrits et traduits. Lorsque le segment de type diversité est présent, la recombinaison D-J est d'abord réalisée (Figure 8 A.B.), suivie de la recombinaison V-DJ (Figure 8 B.C). Considérant la partie variable d'une chaîne de récepteur antigénique, les deux premières régions hypervariables, CDR1 et CDR2 sont codées par le gène V, la région CDR3 est codée par la jonction V(D)J et la région constante de ces chaînes correspond un gène C (pas de région constante pour les deux types de chaînes légères d'immunoglobulines). Le nombre de gènes de chaque type rendus disponibles pour le réarrangement est à l'origine d'une diversité combinatoire, qui constitue un facteur de variabilité pour la jonction V(D)J et donc pour la séquence de la boucle CDR3. La région CDR3 étant responsable d'un grand nombre des contacts directs avec le peptide antigénique, l'immunocompétence d'un individu repose sur la diversité de séquence de la boucle CDR3.

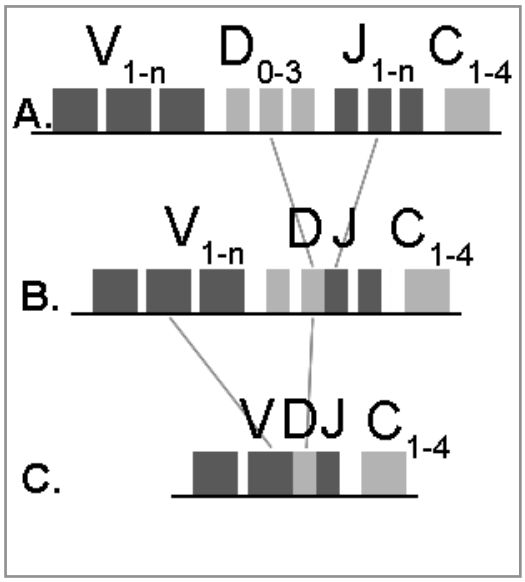


Figure 8. A. Représentation schématique d'un locus de chaîne de récepteur antigénique à l'état germinal. **B.** Recombinaison D-J. **C.** Recombinaison V-DJ.

Les Séquences RSS

La spécificité de site du mécanisme de recombinaison V(D)J repose sur la présence de Séquences Signal de Recombinaison (RSS) disposées de manière contiguë aux portions codantes des gènes V, D ou J. Les gènes V et J possèdent chacun un RSS à une extrémité, alors que les gènes D sont flanqués d'un RSS sur leurs deux extrémités. Les RSS sont les substrats géniques de la recombinase V(D)J (complexe protéique responsable des réarrangements) : ces séquences signal sont toutes constituées d'un heptamètre, d'une séquence espaceur et d'un nonamère (Figure 9.A). L'heptamètre présente une séquence consensus CACAGTG hautement conservée ; le nonamère, de séquence consensus ACAAAAACC, est moins conservé [Glusman *et al*, 2001]. La séquence espaceur peut présenter quelques préférences de séquences du coté du nonamère, mais se trouve par ailleurs non conservée [Ramsden *et al*, 1994]. Les séquences espaceurs ont une longueur de 12±1pb ou de 23±1pb, ce qui définit deux types de RSS, notés RSS-12 et RSS-23, ou également RSS de type 1 et 2 (*one turn RSS* et *two turn RSS*) (Figure 9). L'efficacité de la recombinaison requiert un RSS12 et un RSS23, une restriction nommée règle 12/23 [Tonegawa, 1983]. Dans le cas d'une recombinaison VJ ou d'une recombinaison V(D)J, les gènes à réarranger sont flanqués de RSS de types différents, selon une disposition asymétrique (Figure 9.B). Si la portion codante des gènes V est toujours flanquée d'un RSS de type 2, dans le cas d'une recombinaison VJ, chaque gène J présente un RSS de type 1, alors que, dans le cas d'une recombinaison V(D)J, le gène D présente un RSS de type 1 à chacune de ses extrémités et les gènes J présentent un RSS de type 2. Ces asymétries empêchent les gènes de même type de se réarranger entre eux, seuls des réarrangements V-J, V-D ou D-J pouvant être opérés.

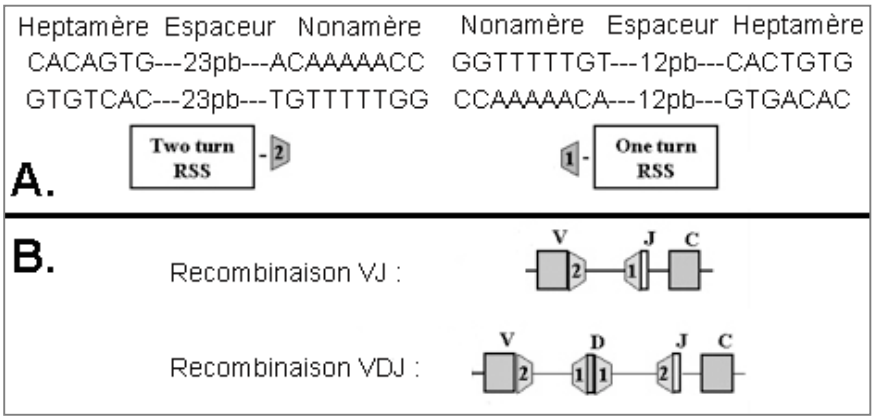


Figure 9. Les Séquences Signal de Recombinaison (RSS). A. Séquences consensus pour l'heptamère et le nonamère et définition des RSS de type un (*one turn RSS*) et 2 (*two turn RSS*) sur la base le la longueur de leur séquence espaceur. **B.** Disposition des RSS de type 1 et 2 pour chaque type de gène V, D, J dans le cas d'une recombinaison VJ et d'une recombinaison V(D)J.

Complexe synaptique et modèle de capture

Lors de l'occurence d'un réarrangement V(D)J, les séquences RSS sont reconnues, rassemblées et maintenues à proximité par le complexe RAG1/2 (*Recombination Activating Gene*), complexe endonucléasique spécifique des lignées lymphoides. Cet ensemble (faisant intervenir également d'autres protéines détaillées par la suite) constitue le complexe synaptique. L'assemblage d'un complexe synaptique obéirait à un modèle de capture : un complexe formé d'un RSS et de la RAG selon une stochiométrie SC2 (un dimère de RAG1 et deux molécules RAG2) serait initialement constitué, et la capture subséquente du RSS partenaire achèverait l'élaboration du complexe PC (*Paired Complex*), (Figure 10.D) [Mundy *et al*, 2002]. Des expériences biochimiques ont ainsi démontré que l'activité de clivage de la RAG était plus importante, lorsque les complexes synaptiques étaient assemblés de manière progressive, en ajoutant des RSS-23 libres à des complexes RAG-RSS-12 (ou vice-versa), que lorsqu'une mise en présence de complexes RAG-RSS-12 et RAG-RSS-23 était réalisée [Jones *et al*, 2002]. A l'intérieur du complexe PC, les RSS adopteraient probablement une configuration en orientation courbée et croisée [Ciubotaru *et al*, 2007].

La phase de clivage de la recombinaison V(D)J

Le réarrangement de deux gènes se divise conceptuellement en une phase de clivage et une phase de jointure (Fig.10.A, B et C). La RAG catalyse la coupure double brin de l'ADN à la jonction RSS/portion codante des deux gènes à réarranger [Gellert *et al*, 2002; Fugmann *et al*, 2000; Schatz and Baltimore, 1988]. Ce complexe introduit initialement une coupure simple brin

à l'extrémité 5' de l'heptamère du RSS, exposant, à chaque extrémité codante, un groupe 3'-hydroxyl (OH), (Figure 10.E). Ce groupe dégagé est ensuite utilisé par la RAG pour catalyser une réaction directe de transestérification sur le pont phosphodiester opposé [McBlane *et al*, 1995], ce qui crée une boucle covalente (structure en épingle à cheveux) à l'extrémité codante et une extrémité signal phosphorylée [Schlissel *et al*, 1993; Roth *et al*, 1992]. Les réactions de transestérification se produisent simultanément au niveau des deux RSS complémentaires, à l'intérieur du complexe synaptique, deux coupures doubles brins couplées sont donc réalisées.

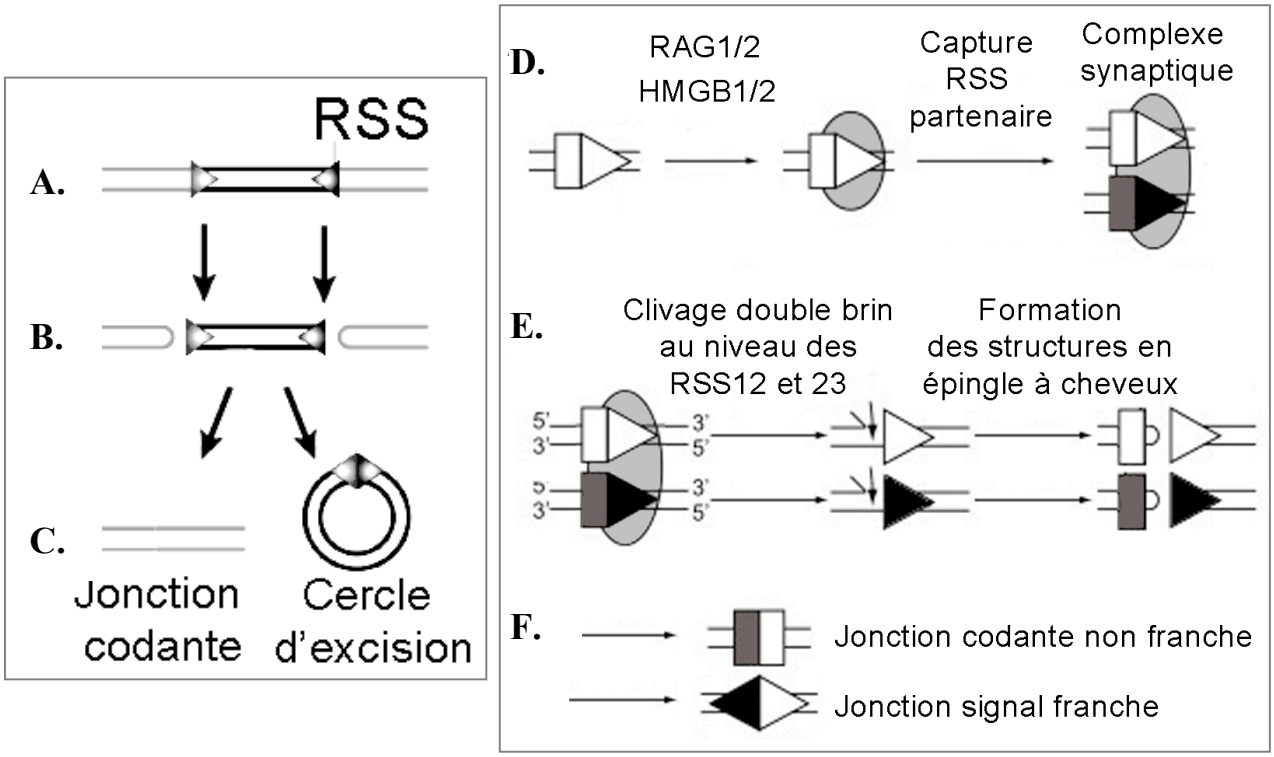


Figure 10. A. Position des deux portions codantes à réarranger, séparées par leurs RSS respectifs et une séquence intermédiaire. **B.** Coupure de l'ADN. **C**. Jointure non franche à la jonction codante et formation du cercle d'excision. **D.** Positionnement du complexe RAG 1/2 sur son RSS cible, capture d'un RSS partenaire et formation du complexe synaptique. **E.** Clivage simple brin et dégagement du groupe 3'-OH, transestérification du second brin et formation des extrémités franches et en épingles à cheveux. **F.** Résolution du réarrangement : formation d'une jonction codante non franche et d'une jonction signal franche.

La phase de jointure de la recombinaison V(D)J

Lors de la phase de jointure de la recombinaison V(D)J, les deux extrémités des portions signal sont liées de manière franche, formant la jointure signal ; la portion chromosomique qui séparait les gènes réarrangés forme un cercle d'excision et se trouve soustraite (Fig.10.C). Les portions codantes sont sujettes à des réactions qui résolvent la structure en épingle à cheveux de manière asymétrique, ce qui peut entraîner une délétion de nucléotides ou former des séquences palindromiques au niveau de la jonction codante. Dans les portions terminales, la désoxynucléotidyl transférase et des DNA polymérases vont ensuite ajouter des nucléotides N au

hasard. Les extrémités codantes sont subséquemment jointes par une machinerie de réparation non-homologue de l'ADN. Le caractère non franc de la jonction codante augmente profusément la variabilité de la séquence V(D)J, créant une diversité jonctionnelle. Au niveau de l'ADN, une séquence intronique séparant le gène J réarrangé et le segment Constant reste présente : elle est transcrite et excisée au niveau de l'ARN (permettant les permutations isotypiques des immunoglobulines).

Réarrangements productifs et non-productifs

La nature non franche de la jonction codante peut entraîner un décalage du cadre de lecture transcriptionel. Du fait de la lecture ribosomale par codons, les réarrangements V(D)J ne préservent la phase ouverte du cadre de lecture que dans 1/3 des cas (*in-frame rearrangements*). Les réarrangements ne pouvant conduire à la synthèse de la chaîne de récepteur antigéniques sont dits non-productifs (*out-of-frame*). La possibilité de genèse d'un codon stop relative à l'imprécision de jonction ainsi que des défauts dans la phase de jointure des réarrangements (menant par exemple à des translocations) diminuent légèrement la proportion de 1/3 de réarrangements productifs [Coleclough, 1983].

Les facteurs impliqués dans la recombinaison V(D)J

La recombinase V(D)J réfère à l'ensemble des enzymes, spécifiques des lymphocytes ou ubiquitaires, responsables de la recombinaison V(D)J. Elle fait tout d'abord intervenir le complexe d'endonucléases RAG1/2, enzymes exprimées à des étapes spécifiques du développement des Lymphocytes B et T. La coupure double brin de l'ADN, catalysée par la RAG, entraîne la fixation d'une protéine ubiquitaire de réparation, la DNA-PK (Protéine Kinase ADN dépendante), qui va recruter plusieurs autres protéines, et notamment la nucléase ARTEMIS ,avec laquelle elle agit de concert pour ouvrir la structure en épingle à cheveux. La genèse efficace d'une jonction codante et signal nécessite l'intervention de la machinerie de réparation des extrémités non-homologues de l'ADN (NHEJ, *nonhomologus end-joining repair pathway*), faisant intervenir la DNA-PKcs, Ku70, Ku80, l'ADN ligase IV, XRCC4, XLF et Cernunnos [Gellert, 2002; Buck *et al*, 2006]. Plusieurs ADN polymérases participent également à la résolution du réarrangement.

RAG1 et RAG2

Les enzymes RAG1 et RAG2 (*Recombination Activating Genes*) sont deux endonucléases codées par deux gènes contigus chez les vertébrés mandibulés, dont l'expression est spécifique des cellules lymphoïdes [Schatz, 1989; Oettinger, 1990]. La délétion d'une ou de l'autre des deux enzymes RAG est suffisante pour inhiber totalement les réarrangements, démontrant leur importance lors de la recombinaison [Shinkai et al. 1992]. Des études d'extraits cellulaires ont établi que les RAG 1 et 2 étaient suffisantes pour permettre le clivage, que ce dernier était concerté au niveau des deux RSS, requérant la structure de synapse et qu'enfin, des mutations au niveau d'un RSS empêchaient le clivage au niveau des deux RSS [McBlane *et al*, 1995; Eastman *et al*, 1996]. Le complexe RAG marque l'initiation d'une recombinaison en se liant à une séquence RSS, très probablement avec l'intervention des protéines HMGB1 et HMGB2. Le complexe RAG reconnaît et se fixe spécifiquement au nonamère et à l'heptamère du RSS [Swanson, 2002]. Plus précisément, des mutations du nonamère, de l'espaceur ou de la partie 3' de l'heptamère n'altèrent que modérément la fixation de la RAG, alors qu'une mutation au niveau des 3 premières bases CAC de l'heptamère entrave très fortement la reconnaissance par la RAG [Akamatsu, et al 1994 ; Ramsden et al, 1994]. L'ajout ou la délétion de plus d'une paire de base au niveau des espaceurs 12 ou 23 réduit aussi très fortement cette reconnaissance. Une fois lié à un RSS, le complexe RAG1/2 HMGB1/2 agit selon un modèle de capture, en

attirant un second complexe RSS/RAG, pour former le complexe synaptique [Jones and Gellert, 2002 ; Mundy et al. 2002]. Les enzymes RAG1/2 et HMGB1/2 se maintiennent au niveau des brins codant et signal postérieurement au clivage double brin ; des protéines de réparation de l'ADN sont alors recrutées par le complexe.

Il est à noter qu'au sein du complexe RAG, la protéine RAG1 joue un rôle direct dans la fixation au RSS et dans la coupure de l'ADN. La RAG1, constituée de 1040 résidus d'acides aminés, contient des domaines pouvant interagir avec le nonamère (NBD *Nonamer Binding-Domain*, domaine amino terminal, résidus 761-979) et l'heptamère (domaine central, résidus 528-760). Ce dernier domaine contient 3 acides aminés (asp-600, asp-708 et glu-963) essentiels pour l'étape de coupure simple brin [De and Rodgers, 2004 ; Fugmann et al, 2000 ; Kim et al, 1999]. Cette triade DDE est d'ailleurs un motif retrouvé dans de nombreuses transposases et intégrases [Haren *et al*, 1999], et contribue à un clivage du double brin en trans, lors du réarrangement [Swanson, 2001]. Le domaine C-terminal de RAG1 (résidus 761-979) est responsable d'une liaison à l'ADN double brin non-spécifique et coopérative. Les fonctions RAG2 (517 a.a.) sont moins bien déterminées. Cette protéine ne présente pas d'activité de fixation à l'ADN par elle-même, mais RAG2 interagit avec RAG1, pour améliorer la spécificité et l'affinité de la fixation aux RSS [Swanson *et al*, 2004]. RAG2 augmente également l'activité de clivage du complexe RAG in vitro [Shimazaki et al., 2009].

HMGB1 et HMGB2

Le complexe RAG nécessite la présence des enzymes HMGB1 et HMGB2 (*High Mobility Group-Box family*), pour assembler efficacement une paire de RSS et réaliser les clivages double brin en respectant la règle 12/23 [van Gent *et al*, 1997]. De plus, une sur-expression de HMGB1/2 augmente l'occurrence des réarrangements V(D)J [Aidinis et al, 1999]. HMGB1 et HMGB2 ne sont pas spécifiques de la recombinaison V(D)J : comme les autres protéines du groupe HMG (*High Mobility Group*), elles constituent des protéines non-histoniques associées à l'ADN chromosomique. Ubiquitairement distribuées dans les noyaux des cellules eukaryotes, elles collaborent à la transcription, réplication, recombinaison et réparation de l'ADN, en soutenant l'assemblage de complexes nucléoprotéiques [Thomas, 2001]. Les protéines HMGB1 et HMGB2 (*High Mobility Group Box-domain*) facilitent les interactions coopératives entre des protéines d'activité cis, en augmentant la flexibilité de l'ADN et en permettant de ce fait l'assemblage de gros complexe protéiques liés à l'ADN. La fixation de ces protéines se réalise au niveau du petit sillon de l'ADN, cette affinité structurelle étant indépendante de la séquence nucléotidique. En tant que constituantes du complexe synaptique de

recombinaison, ces protéines jouent un rôle double, en amenant des éléments critiques de l'heptamère du RSS-23 dans la meme phase que le RSS-12, promouvant ainsi la fixation de la RAG, et également en assistant la catalyse du clivage au niveau du RSS-23 [Swanson *et al*, 2002]. C'est précisément cette amélioration de la présentation des RSS au complexe RAG, permise par la distortion structurale de la séquence espaceur qui stabilise l'interaction RAG/ADN, accroîssant le rendement de coupures double brin entre RSS et portions codantes [Yoshida et al, 2000].

DNA-PK

La DNA-PK est une protéine kinase serine/thréonine dépendante de l'ADN, constituée d'une sous-unité catalytique DNA-PKcs et des antigènes auto-immuns Ku70 et Ku80. Cette protéine intervient de façon ubiquitaire dans la ligation non-homologue des coupures double brin de l'ADN (NHEJ). Si chacune des 3 sous-unités peut fixer l'ADN [Yaneva et al. 1997], néanmoins, le site catalytique seul est inactif. L'activité de la DNA-PK repose sur les hélicases Ku, qui dirigent la protéine vers les cassures double brin de l'ADN et ciblent l'activité kinase. Lors d'un événement de recombinaison V(D)J, la DNA-PK intervient lors de l'étape de résolution du réarrangement. Une fois la protéine fixée a l'ADN, la sous-unité catalytique DNA-PKcs s'autophosphoryle, ce qui induit un changement de conformation permettant l'accès des enzymes de réparation aux extrémités dégagées par la coupure double brin [Meek, et al. 2008] : ARTEMIS et les autres protéines de la recombinase V(D)J sont ainsi recrutées.

ARTEMIS

La protéine ARTEMIS est une nucléase qui intervient dans la recombinaison V(D)J, en permettant la réouverture asymétrique des structures en épingle à cheveux présentes aux extrémités codantes. Seule, cette enzyme présente une activité exonucléasique simple brin 5'-3' [Ma et al. 2002]. Ce n'est que complexée avec la DNA-PKcs et phosphorylée, qu'ARTEMIS acquiert son activité endonucléasique, permettant l'ouverture asymétrique des structures en boucles covalentes des extrémités codantes [Pannicke et al. 2004].

XRCC4 et Cernunnos

Les protéines XRCC4 (*X-ray Repair Cross-Complementing factor 4*) et Cernunnos (également nommé XLF ou XRCC4-*like factor*) sont des protéines de réparation de l'ADN, qui agissent en concert avec la DNA-PK pour aligner deux extrémités d'ADN. La délétion de XRCC4 inhibe la

résolution des extrémités signales et codantes. XRCC4 et Cernunnos participent également au recrutement de la TdT.

TdT ou DNTT

La terminale transférase, nommée TdT (*Terminal deoxynucleotidyl Transferase*) ou DNTT (*DNA nucleotidylexotransferase*) est une DNA polymérase spécialisée, qui ajoute N-nucléotides aux exons V, D ou J, durant la recombinaison des loci de récepteurs antigéniques, avant que la coupure ne soit jointe [Gilfillan et al, 1993 ; Komori et al, 1993]. Contrairement à la plupart des ADN polymérases, la TdT ne nécessite pas un modèle complémentaire pour l'ajout de nucléotides ; la séquence ajoutée est aléatoire et diversifie très fortement la jonction codante.

ADN polymérases μ et/ou λ

A la suite de l'addition de nucléotides par la TdT, les ADN polymérases λ and μ insèrent les nucléotides additionnels nécessaires pour rendre les deux extrémités codantes compatibles, en vue de la liaison [Bertocci *et al*, 2006].

La ligase IV et XRCC4

La ligase IV relie les extrémités libres de l'ADN, complétant ainsi la phase de jointure des réarrangements [van Gent and van der Burg, 2007]. son association avec XRCC4, permise par deux motifs enzymatiques de 5 acides aminés (BRCAT), favorise cette activité de résolution des réarrangements.

ATM et p53

Les protéines ATM (*Ataxia Telangiectasia Mutated*) et p53 (*tumor protein* 53; 53 pour 53kDa) sont recrutées et intègrent la recombinasse V(D)J, avant la résolution des coupures doubles brins. La protéine ATM occupe une place centrale dans la signalisation des dommages de l'ADN, lors des points de vérification du cycle cellulaire. La protéine ubiquitaire p53, quant à elle, est de grande importance pour les organismes multicellulaires : surnommée la gardienne du génome, elle régule le cycle cellulaire et conserve la stabilité du génome, en prévenant les mutations [Read and Strachan, 1999]. La recombinaison V(D)J incluant des coupures double brins de l'ADN, cette activité recombinasse se doit d'etre contenue exclusivement à l'intérieur des phases non proliférantes du cycle cellulaire. Les thymocytes ATM(-/-) et p53(-/-) présentent une persistance des coupures double brin à travers le cycle cellulaire. L'intervention la voie de

signalisation ATM/p53, au cours des mécanismes de recombinaison V(D)J, est donc essentielle pour contenir les coupures de l'ADN dans des cellules non proliférantes [Dujka et al, 2009]. Cette restriction est cruciale pour préserver l'intégrité du génome des lymphocytes en développement [Kang *et al*, 2010].

Contrôle des Réarrangements V(D)J

Les réarrangements V(D)J sont opérés au niveau des différents loci de récepteurs antigéniques, exclusivement à l'intérieur des lignées lymphocytaires et à des stades très précis de leur développement [Yancopoulos and Alt, 1985 ; Cobb et al., 2006]. Ces réarrangements sont donc locus spécifiques et temporellement régulés. La réalisation d'une telle prouesse en matière de régulation repose sur un ensemble de mécanismes incluant une activité promotrice différentielle des enzymes RAG1/2, ainsi qu'un remodelage chromatinien, en lien avec l'activité transcriptionnelle germinale non productive et avec la position des loci à l'intérieur du noyau cellulaire. Bien que complexes, ces contrôles peuvent être perçus comme une régulation de nature substrat/enzyme conduisant à la mise en place de centres de recombinaison au niveau des loci cibles, durant les stades de développement lymphocytaires appropriés.

Activité promotrice différentielle des enzymes RAG1/2

L'expression conjointe des enzymes RAG1 et RAG2 est limitée aux lymphocytes en cours de développement. L'activité promotrice de l'enzyme RAG2 est différentielle entre les lignées lymphocytaires B et T : la région promotrice est située à -7Kb en amont du gène dans les cellules B, et, pour les cellules T, cette séquence s'étend sur une région allant de –2Kb à –7Kb [Monroe et al. 1999]. Si les variations de séquence des RSS n'affectent que relativement peu la fréquence d'utilisation des gènes correspondant [Feeney *et al*, 2000; Wilson et al, 2001], ces séquences ne rendent en aucun cas compte des changements dynamiques de spécificité de la recombinase V(D)J, pour les différents loci qui accompagnent les différentes étapes de différentiation. Au cours du développement lymphocytaire, la propension d'un locus de chaîne de récepteur antigénique à être réarrangé serait déterminée par l'accessibilité qu'il offre au complexe RAG. Cette vision attribue un rôle plutôt passif à la recombinase V(D)J, les modifications de la structure chromatinienne se plaçant en position prédominante dans la régulation des réarrangements.

La structure chromatinienne

La chromatine peut se trouver schématiquement sous un état condensé ou décondensé. L'hétérochromatine, formée de fibres de 200 à 300 angströms de diamètre, correspond à de la chromatine à l'état condensé. Sous cet état, l'ADN se trouve compacté autour d'octamères de

protéines d'histones, de structure 2*(H2A, H2B, H3, H4). Environ 146 paires de bases s'enroulent autour de chaque corps octamérique pour former un nucléosome, qui constitue la première compaction de l'ADN. Une seconde étape de compaction produit la fibre de 30 nm. Les nucléosomes empêchent physiquement l'accès à l'ADN : lorsqu'une séquence d'ADN est située dans une structure nucléosomale, la capacité d'un facteur à s'y fixer est fortement inhibée [Wolffe and Guschin 2000]. Les contacts histone/ADN peuvent être perturbés par des modifications post-traductionnelles des extrémités amino-terminales ou queues histoniques, qui se projettent à l'extérieur de la partie globulaire des histones. Ces modifications post-traductionnelles comprennent des réactions de méthylation ou d'acétylation et sont catalysées par des enzymes spécifiques (histone-acétyltransférase, histone-déacétylase, histone-méthylase, histone-kinase).

L'acétylation des histones est produite par les histones-acétyl-transférases (HAT). Ces protéines catalysent le transfert d'un groupement acétylé de charge négative sur un résidu histonique lysine, acide aminé chargé positivement. La neutralisation de la charge positive de la queue histonique provoque la répulsion des charges négatives des phosphates de l'ADN et déstabilise le nucléosome [Hong et al. 1993; Gorish et al. 2005]. Ainsi, l'acétylation au niveau de l'histone H4K16 suffit à empêcher la compaction de la chromatine formant la fibre de 30 nm [Shogren-Knaak et al. 2006]. L'acétylation des histones, en particulier des H3 et H4, réduit leurs charges positives et participe à la désorganisation des nucléosomes.

La méthylation des histones peut s'effectuer sur les résidus lysine ou arginine. Il s'agit d'une modification post traductionnelle assez stable ; jusqu'alors, très peu d'histones déméthylases ont été découvertes. La méthylation des histones peut favoriser ou empêcher la compaction de la chromatine, en fonction des résidus méthylés.

La méthylation de l'ADN est une autre marque épigénétique, qui a pour conséquence de diminuer la capacité d'un facteur à se fixer sur sa séquence cible au niveau de l'ADN. Les cytosines 5-méthyltransférases sont les enzymes responsables du transfert de groupements méthyle sur les résidus cytosine de l'ADN. Au niveau de l'ADN, seules les bases cytosine (C) peuvent être méthylées dans les îlots CpG. Schématiquement l'état d'hétérochromatine correspond à des histones hypoacétylés et à de l'ADN hyperméthylé par rapport à la chromatine décondensée ou euchromatine.

Il est à noter qu'il existe une relation entre le niveau de méthylation de l'ADN et le niveau d'acétylation des histones. L'ADN méthylé provoque le recrutement d'histones déacétylases et, inversement, l'acétylation des histones peut affecter la méthylation de l'ADN [Jones et al. 1998]. Cet antagonisme entraîne la mise en place d'un certain équilibre dynamique

entre les territoires hétérochromatiniens et euchromatiniens.

Des complexes de remodelage des nucléosomes, comme SWI/SNF, ont pour fonction d'augmenter l'accessibilité des facteurs protéiques à l'ADN nucléosomal, par un processus ATP-dépendant de déplacement des octamères d'histones le long de l'ADN ou de perturbation locale des contacts histone/ADN [Narlikar, et al. 2002; Becker and Horz, 2002]. Ces complexes de remodelage de l'ADN peuvent être ciblés sur des promoteurs, via leur association avec des facteurs protéiques, comme les activateurs de la transcription.

Remodelage de la chromatine et contrôle des réarrangements V(D)J

Concernant l'activité des réarrangements V(D)J, en lien avec le remodelage chromatinien, il est reconnu que les nucléosomes inhibent l'initiation de la recombinaison V(D)J, en empêchant RAG1 et RAG2 de se lier à leurs séquences RSS cibles [Golding et al. 1999; Kwon et al. 1998; McBlane and Boyes, 2000]. Les séquences RSS elles-mêmes répriment l'accessibilité, en causant le positionnement préférentiel du nucléosome sur le RSS. Ce positionnement est en effet entraîné *in vitro* et *in vivo* par la séquence nonamère conservée du RSS [Baumann et al. 2003]. Le remodelage de la chromatine, préliminaire aux réarrangements, peut impliquer des complexes de remodelage de l'ADN : ainsi, SWI/SNF facilite la coupure par la RAG au niveau de structures mono-nucléosomiques *in vitro* [Kwon J, Morshead et al. 2000].

Acétylation des Histones

Durant la recombinaison V(D)J, la chromatine doit être remodelée, non au niveau d'une région promotrice spécifique, mais à différents points à l'intérieur du locus de récepteur antigénique à réarranger. De nombreuses observations ont permis de vérifier le rôle activateur de l'acétylation sur les réarrangements.

Les loci en phase de réarrangement V(D)J présentent une augmentation de leur niveau d'acétylation [McMurry and Krangel, 2000; Roth et al. 2000; Huang et al. 2001; Ye et al, 2001]. L'acétylation des histones stimule la recombinaison *in vivo*, alors que l'inhibition des enzymes histones déacétylases induit une augmentation des réarrangements *in vitro* [McBlane and Boyes, 2000]. L'hyperacétylation des histones H3 sur le mini-locus TCRδ, ainsi que sur le locus TRAD endogène, est corrélée positivement avec le niveau de recombinaison V(D)J. Concernant le mode d'action, l'acétylation augmenterait l'activité des complexes de remodelage de l'ADN, en favorisant leur accessibilité aux nucléosomes. L'acétylation et les complexes de remodelage

agiraient ainsi en synergie pour augmenter l'activité de recombinaison au sein d'un locus [Nightingale et al. 2007].

Méthylation de l'histone H3K4

Une autre modification post-traductionnelle des histones, la triméthylation de l'histone H3 sur son résidu lysine 4 (H3K4me3), joue un rôle majeur dans l'induction des recombinaisons. La diméthylation et la triméthylation de l'histone H3K4 augmentent les réarrangements D-J au niveau du locus IgH [Chakraborty *et al*, 2007; Liu *et al*, 2007; Matthews *et al*, 2007; Morshead *et al*, 2003]. Un domaine PHD (*Plant HomeoDomain*) de RAG2 a été identifié : ce groupement fixerait spécifiquement l'histone H3, lorsqu'il porte une triméthylation [Liu *et al*, 2007; Matthews *et al*, 2007].

Méthylation de l'histone H3K9

La diméthylation de l'histone H3 sur son résidu lysine 9 (H3K9me2), associée à la chromatine silencieuse (non transcrite), est corrélée avec une inhibition de la recombinaison V(D)J [Johnson *et al*, 2004 ; Osipovich *et al*, 2004]. Le groupement diméthyle est retiré des histones H3K9 au niveau des nucléosomes situés sur la région V_H, et ce, spécifiquement durant le stade d'occurrence des réarrangements V_H-D_H-J_H [Johnson *et al*, 2004].

Vers une intégration des signaux épigénétiques contrôlant la recombinaison V(D)J

Si la mise en évidence de la reconnaissance de l'histone H3K4me3 par le domaine PHD de RAG2 [Liu *et al*, 2007; Matthews *et al*, 2007] constitue un premier pont entre les modifications épigénétiques des loci et leur activité de réarrangement, il est évident que cette marque épigénétique est trop largement distribuée au niveau du génome global, pour agir seule dans le ciblage de la recombinase V(D)J. Même s'il semble évident que ce ciblage doive impliquer nécessairement une somme combinatoire de modifications chromatiniennes, la réponse est loin d'être élucidée. La question du mécanisme reliant les signaux de différentiation et le dépôt ou le retrait de ces marques épigénétiques reste également ouverte [Liu *et al*, 2009].

Transcription ARN non-codante (Non coding RNA transcription)

Contrairement au dogme selon lequel l'ADN donne l'ARN et l'ARN la protéine, de nombreuses analyses de la transcription au niveau du génome global estiment que plus de la moitié des séquences transcrites chez les mammifères ne codent pas pour des protéines [Mattick

et al, 2005], cette transcription non-codante présentant un rôle régulateur des gènes. Concernant la recombinaison V(D)J, le locus IgH est le premier locus pour lequel une transcription non-codante, stérile ou germinale ait été identifiée. Cette transcription germinale est initiée à partir de l'enhancer Eμ et du promoteur PDQ52, immédiatement en amont des segments D_H, juste avant la recombinaison D_H-J_H [Lennon *et al*, 1985 ; Thompson *et al*, 1995]. Puis la transcription germinale est initiée à partir des segments V_H, et la recombinaison V_H-DJ est réalisée [Yacopoulos *et al*, 1985 ; Corcoran *et al*, 1998]. La question d'une transcription germinale résultant de l'accessibilité d'un locus ou du rôle actif de la transcription dans l'acquisition de cette accessibilité a alors été soulevé. Les gènes V, D ou J sont relativement courts et sont séparés par de nombreuses séquences introniques. Aussi, une transcription germinale génique et intergénique du locus est présente à chaque étape de réarrangement D-J et V-DJ du locus IgH et disparaît lorsque le locus a achevé son réarrangement [Bolland *et al*, 2004]. L'identification d'une transcription anti-sens non contrôlée par les promoteurs V_H est la première évidence d'un rôle fonctionnel de la transcription germinale dans l'activation des réarrangements V(D)J au niveau du locus IgH. Cette transcription non codante pourrait remodeler la région V_H, probablement en ciblant l'action de facteurs de remodelage de la chromatine. La présence d'une transcription non-codante anti-sens stimulatrice de l'accessibilité a été également observée pour d'autres loci [Krangel, 2003]. Plus généralement, il a été décrit postérieurement que la transcription d'un RSS par la polymérase II présentait un rôle essentiel dans le réarrangement du gène correspondant [Abarrategui and Krangel, 2006]. La polymérase II est donc fortement impliquée dans la déstabilisation des nucléosomes, facilitant l'accès à la recombinase V(D)J.

Repositionnement subnucléaire (Subnuclear relocalisation)

La localisation physique d'un locus de chaîne de récepteur antigénique dans le noyau est déterminante pour son potentiel de recombinaison. Dans les cellules B immatures procédant aux recombinaisons des loci IgH et IgK, les allèles actifs pour les réarrangements sont repositionnées vers l'intérieur euchromatique du noyau, région permissive à la transcription (l'hétérochromatine est localisée principalement en périphérie du noyau et du nucléole, alors que l'euchromatine est répartie à l'intérieur du nucléoplasme) [Kosak *et al*, 2002]. Cette relocalisation est dépendante du signal du récepteur à l'interleukine Il-7. Ce repositionnement intra-nucléaire a également été constaté lors des réarrangements des loci TCRA et TCRB [Kosak *et al*, 2002].

Boucle Chromatinienne

Si le repositionnement allélique du locus IgH est suffisant pour les réarrangements D_H-J_H et les réarrangements V_H-D_H-J_H impliquant les segments géniques V_H proximaux du locus, il ne permet pas les recombinaisons faisant participer les gènes V_H proximaux, probablement du fait de la grande taille du locus. Pour ces réarrangements, un procédé supplémentaire, nommé concentration du locus, est nécessaire : il va juxtaposer les gènes V_H distaux des gènes D_H-J_H réarrangés via des boucles chromatiniennes engageant les sous-domaines du locus IgH [Roldan *et al*, 2005 ; Sayegh *et al*, 2005]. Cette contraction en boucles a également été observée dans le cas des réarrangements des loci TRA et TRB [Skok *et al*, 2007].

Interactions Promoteurs-Enhancers et Accessibilité Chromatinienne

Chacun des 7 loci codant pour les récepteurs antigéniques présente des éléments de contrôle en Cis, de type enhancers ou promoteurs, qui ciblent les enzymes de modification et de remodelage de la chromatine, ainsi que la polymérase II. Ces enzymes agissent conjointement, pour rendre l'ADN accessible et activer la fixation de la RAG.

Enhancers et usines de transcription

Dans le noyau, la transcription ne s'opère pas de manière homogène, mais se trouve concentrée dans des régions subnucléaires, les usines de transcription, contenant chacune probablement jusqu'à dix ARN polymérases II et pouvant transcrire plusieurs gènes simultanément [Faro-Trindade *et al*, 2006]. Les enhancers peuvent relocaliser les loci de récepteurs antigéniques auxquels ils sont associés, depuis la périphérie nucléaire jusqu'aux usines de transcription [Ragoczy *et al*, 2006].

Centres de recombinaison

La fixation de la RAG, fortement régulée durant le développement des lymphocytes, se réduirait a une courte région comprenant les gènes J (et, lorsque les segments D sont présents, la région J-gènes D proximaux) dans les loci IgH, IgK, TCRB, and TCRA. Ces régions, définies comme centres de recombinaison, sont riches en modifications histoniques activatrices (Acétylation en H3, H3K4me3) et en polymérase II [Jung et al., 2006]. Elles présentent de ce fait toutes les caractéristiques favorables au recrutement et à la fixation du complexe RAG. Alors que la fixation de RAG2 serait globale dans le génome au niveau de sites présentant des niveaux élevés de H3K4me3, la fixation de RAG1, plus réduite, car requérant une interaction directe avec

le RSS, pourrait expliquer un recrutement préférentiel du complexe RAG au niveau des centres de recombinaison [Ji et al. 2010]. Ce recrutement serait alors indépendant du type de RSS 12/23 et cette première fixation de la RAG amorcerait la capture du second RSS, pour former le complexe synaptique.

Réarrangement des loci TRA et TRB : exclusion/inclusion allélique, sélections thymiques et biais de répertoires

Au cours du développement des lymphocytes T, la fonctionnalité des TCR générés est vérifiée, de manière à garantir que, seuls, les lymphocytes fonctionnels et non-autoréactifs migreront à la périphérie de l'organisme. Les sélections thymiques permettent de préserver l'intégrité de l'organisme face aux flux constants de nouveaux lymphocytes, et engendrent colatéralement une certaine évolution du répertoire T, tout au long de la différenciation thymique.

Réarrangement du locus TRB : exclusion allélique et sélection β

Le stade DN3 est l'étape de développement lymphocytaire correspond à un fort taux de réarrangement du locus TRB (cette étape marque la division entre les lignées Tαβ et Tγδ). Pour rappel, si le réarrangement du locus TRB est productif, la chaîne β est synthétisée et s'associe avec la chaîne pTα (chaîne α immature) et le complexe CD3, pour former le pré-TRC, récepteur antigénique immature. Les réarrangements du locus TRB sont soumis à un mécanisme d'exclusion allélique dépendant du pré-TCR : les souris pTα-/- présentent une augmentation du pourcentage de cellules ayant deux allèles β productifs, en comparaison avec les souris normales, qui ne comportent qu'un réarrangement β productif par cellule [Aifantis, Buer et al. 1997]. Le locus TRB est soumis à un mécanisme d'exclusion allélique : un premier allèle est réarrangé, le second allèle ne réarrange que si le premier réarrangement est non-productif et, finalement, les cellules n'ayant pas eu de réarrangement productif sur leur deux chromosomes, ne pouvant pas produire de chaîne β, sont éliminées par une sélection β [Wilson et al, 2001]. La sélection β s'opère durant la transition DN3-DN4, et l'expression du pré-TCR à la surface cellulaire indique le succès de cette première étape de sélection du répertoire TCR. Cette sélection β garantit que le réarrangement à été productif et permet à la cellule de poursuivre sa différenciation en passant au stade DN4. De ce fait, si, au stade DN3, moins de 30% des réarrangements des gènes TCRβ sont en phase, au stade DN4, près de 2/3 des réarrangements quantifiés sont productifs [Dudley, Petrie et al. 1994]. Chez la souris, 95% des cellules expriment une chaîne β intracellulaire au stade DN4, contre 40% chez la souris pTα-/- [Buer, Aifantis et al. 1997]. Finalement, la superposition des mécanismes d'exclusion allélique des réarrangements du locus TRB et de

sélection β au stade DN3 engendrent la production de cellules DP de spécificité antigénique clonale, compétentes pour la synthèse de la chaîne TCRβ.

Association des chaînes TCRβ/pTα: absence de biais de répertoire

L'exclusion allélique, qui s'applique aux réarrangements du TRA au niveau génomique, est régulée par le pré-TCR. De ce fait, un biais du répertoire pourrait émerger, dû à une association préférentielle de certaines chaînes β avec la chaîne immature pTα. L'analyse quantitative des répertoires TCRβ, aux stades DN3 et DN4, montre que le répertoire TCRβ intracellulaire est similaire avant et après sélection β. Ce répertoire s'apparente également à celui quantifié chez la souris pTα-/-, génétiquement incompétente pour la genèse du pré-TCR [Wilson *et al.* 2001]. Ces observations démontrent qu'une association avec une chaîne TCRβ quelconque est suffisante pour passer avec succès la sélection β. La protéine pTα n'induit en conséquence pas de biais dans le répertoire TCRβ précoce.

Réarrangement du locus TRA : inclusion allélique génomique et exclusion allélique phénotypique incomplète

Au cours de la lymphopoïèse, la chaîne alpha est réarrangée majoritairement durant les stades Immature Simple Positif et Double Positif (ISP-DP). Contrairement aux chaînes β, le locus TRA n'est pas soumis au mécanisme d'exclusion allélique au niveau génomique. La plupart des cellules présentent deux allèles TRA réarrangés, et une fraction cellulaire de 20% à 25% de la population LTαβ globale possède deux types de TCR cytoplasmiques distincts, du fait de l'occurrence d'un réarrangement productif sur chacun des allèles du locus TRA (Figure 11.A) [Malissen, 1992 ; Davodeau 2001], pouvant ou non exprimer les deux types de TCR en surface. L'expression de deux types de TCR en surface ne représente, selon les auteurs, que 2 à 4% des cellules, ou une fraction plus importante d'environ 8% de la population LTαβ globale : les deux chaînes de TCRα cytoplasmiques s'apparient alors avec la chaîne TCRβ, ce dont résulte l'expression duelle en surface de deux types de TCRαβ, aux spécificités distinctes (*dual cells*) (Figure 11.B) [Corthay 2001]. La compétition des deux chaînes TCRα pour la chaîne TCRβ et pour l'expression de surface aboutit donc à une exclusion allélique phénotypique TCRα incomplète.

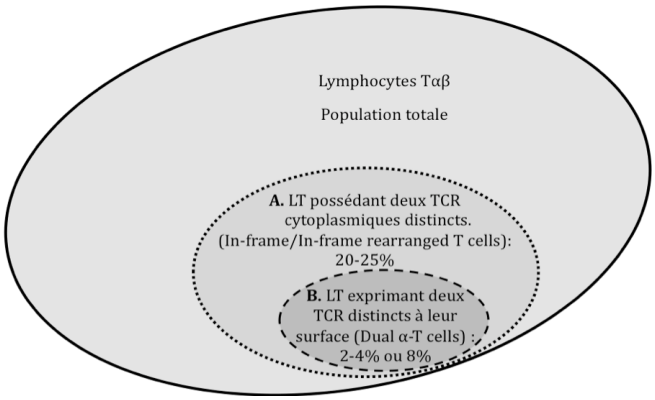


Figure 11. Représentation des cellules LT possédant deux type de TCR cytoplasmiques distincts et des LT exprimant deux types de TCR de spécificité différente en surface, à l'intérieur de la population LTαβ totale. L'ensemble A\B représentent les LT exprimant deux types de TCR cytoplasmiques, mais ayant seulement un type en surface D'après [Malissen, 1992; Munir, 1998, Davodeau 2001 ; Munir, 1998; Corthay 2001].

Association des chaînes TCRα/TCRβ et biais de répertoire

Le biais du répertoire engendré par l'association de la chaîne TCRα avec la chaîne TCRβ serait en faveur des hétérodimères TCRα les plus stables, mais il est difficile d'estimer précisément l'importance de ce biais sur le répertoire global. Néanmoins, un impact de la boucle CDR3 (3ème région hypervariable du domaine variable) sur le répertoire est observé. La boucle CRD3 d'une chaîne de TCR ou d'immunoglobuline présente des variations de longueur dues à l'imprécision de jonction, qui comporte des étapes de délétions et d'additions aléatoires de nucléotides. Contrairement aux associations des chaînes IgH-Igλ et TCRδ-TCRγ, les longueurs de CDR3 des chaînes TCRα et TCRβ retrouvées associées sont très similaires l'une de l'autre [Pannetier et al. 1993; Rock et al. 1994]. Cette association non aléatoire fondée sur la longueur des régions CDR3 induirait un biais dans le répertoire. La taille des boucles CDR3 serait majoritairement contrôlée directement au niveau du mécanisme de régulation des réarrangements, par un mécanisme en amont des étapes de sélection thymique [Hughes et al. 2003].

La sélection thymique

La sélection thymique est réalisée au cours du développement des lymphocytes Tαβ : elle permet de garantir que, seuls, les lymphocytes fonctionnels et non-autoréactifs migreront à la périphérie de l'organisme. C'est au stade Double Positif de leur développement que les lymphocytes T αβ deviennent progressivement capables d'exprimer les TCR à leur surface (TCR^{low}, TCR^{int} et TCR^{high}) [Pearse et al. 1989]. La sélection thymique est basée sur l'affinité des TCR pour les complexes CMH-peptide et se définit par deux seuils d'activation, en fonction de l'avidité de la fixation du récepteur T : le seuil d'efficacité et le seuil d'intégrité (Figure 12) [Naumov *et al*, 2003]. Toutes les cellules nucléées d'un organisme vertébré arborent à leur surface des molécules du complexe majeur d'histocompatibilité uniques (CMH). Au cours de la sélection thymique, c'est l'intensité de l'interaction des TCR exprimés à la surface d'un lymphocyte donné avec les molécules du CMH des cellules épithéliales du cortex thymique et/ou des cellules dendritiques thymiques, qui va déterminer le devenir du lymphocyte considéré. Les 3 destins sont possibles pour les thymocytes, déterminés par un gradient d'activation, fonction de la force et de la durée d'interaction avec les cellules stromales [Alam et al. 1996]. Une trop faible reconnaissance (située en dessous du seuil d'efficacité) indique que le lymphocyte est non fonctionnel, ses TCR de surface étant incapables de se lier efficacement aux molécules du CMH : cette affinité trop faible ne permet pas de générer un signal de survie suffisant pour la cellule, qui enclenche alors un programme de mort cellulaire et meurt par négligence. Une interaction modérée validerait l'efficacité d'un TCR donné à reconnaître le CMH, sans pour autant s'activer contre le peptide du soi. Le lymphocyte T reçoit alors des signaux de survie lui permettant de poursuivre sa différentiation vers le stade Simple Positif (SP) : la cellule est sélectionnée positivement [Sprent et al. 1988; Williams et al. 1996]. Une reconnaissance trop forte (au dessus du seuil d'intégrité) comporte un risque pour l'organisme, cette affinité importante sur-activant le thymocyte et induisant sa mort par précipitation de l'apoptose. Les cellules T effectrices, pouvant réagir contre les cellules du soi et détruire des tissus sains en périphérie, sont alors détruites par un mécanisme de mort par apoptose au cours de la sélection thymique négative.

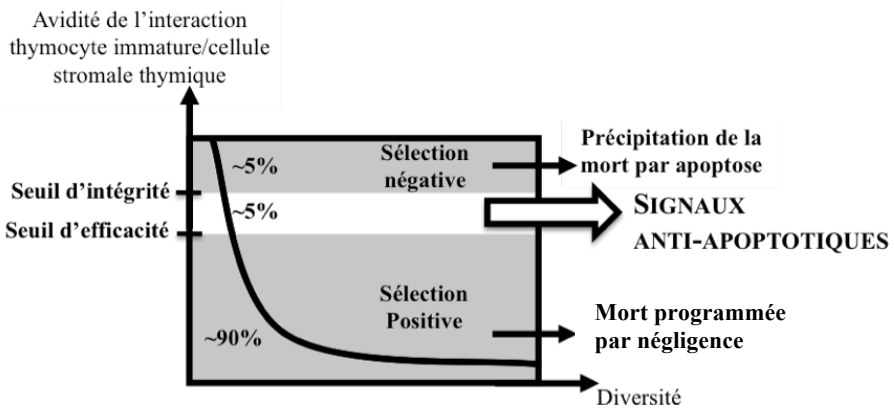


Figure 12. Schématisation de la sélection thymique des lymphocytes Tαb. La diversité relative du répertoire T est représentée en abscisse et l'affinité de l'interaction TCR pour le complexe CMH+peptide du soi figure en ordonnée. Deux seuils d'efficacité du TCR définissent le destin cellulaire des lymphocytes sélectionnés, entre sélection positive, sélection négative et succès du passage de la sélection thymique qui est à l'origine des signaux anti-apoptotiques pour la cellule qui continue alors sa différenciation, passant du stade Double Positif (DP) au stade Simple Positif (SP). Les pourcentages sont indicatifs de la proportion des cellules sélectionnées. Adapté de [Naumov *et al*, 2003].

Plus précisément, la sélection positive s'opère au niveau de l'interface cortico-médullaire thymique, et les cellules DP TCR^{low} qui passent avec succès cette sélection voient leur densité de surface en TCR augmenter. De plus, ces lymphocytes initient l'expression de la molécule CD5, qui peut être vue comme marqueur d'affinité du TCR, sa concentration étant une fonction croissante de l'affinité TCR/CMH-peptide [Bhandoola et al. 1999]. Le passage de la sélection positive correspond à l'expression en surface de la molécule CD69 [Anderson et al. 1999] et à la répression de l'expression génique de RAG1/2.

L'interaction TCR/CMH n'est pas suffisante pour sélectionner les structures efficaces de TCR lors de la sélection positive, et la présence des peptides du soi complexés au CMH contribue nécessairement à l'éducation des thymocytes et à leur survie en périphérie. Des études réalisées sur des animaux transgéniques ne pouvant plus apprêter et présenter les antigènes du soi ont démontré que la mise en place d'un répertoire T fonctionnel nécessitait cette présentation des peptides du soi [Surh et al. 1997; Tourne et al. 1997]. La sélection positive induit un signal de survie pour la cellule DP, lui permettant de passer à l'étape de sélection thymique négative.

La sélection négative se déroulerait essentiellement dans la médulla du thymus et au niveau de la jonction cortico-médullaire, du fait de la migration centripète des cellules DP dans le thymus [Klein and Kyewski 2000]. Les cellules T se trouvant activées par l'association CMH-peptide du soi présentées par les cellules dendritiques de ces compartiments thymiques, sont éliminées ou anergisées [Laufer et al. 1999]. Suite à sa sélection, le niveau d'expression du

récepteur CD5 en surface cellulaire est une fonction croissante du niveau d'affinité du TCR dans son association avec les molécules du CMH-peptide.

Ce modèle de sélection thymique est dépendant de paramètres croisés : un TCR de forte avidité, exprimé en faible densité à la surface cellulaire, peut induire le même signal qu'un TCR de faible affinité exprimé à une plus forte concentration [Viola and Lanzavecchia 1996; Benoist and Mathis 1997]. Concernant la sélection des cellules duelles, qui comportent en surface deux TCR de spécificités distinctes, un seul des deux types de TCR serait impliqué dans le succès du passage de la sélection thymique. Effectivement, l'expression d'une seconde chaîne TCRα en surface ne signifie pas que le second TCR correspondant soit sélectionné : la cellule peut être sélectionnée via un hétérodimère TCRαβ et un échappement de la deuxième combinaison TCRαβ à la sélection pourrait être permise par son expression en faible densité à la surface cellulaire [Health, 1993]. Cette non sélection du second TCR peut mener à la formation de cellules potentiellement multi-réactives. Des études visant à déterminer le rôle de ces cellules duelles dans les maladies auto-immunes ont montré une implication non systématique, observée seulement dans quelques maladies auto-immunes [Corthay 2001, Elliot, 1995].

Pour conclure, chaque individu héritant d'une identité génétique unique au travers des molécules du CMH et de différents haplotypes des gènes V, (D) et J codant pour les récepteurs des cellules T, le système de sélection dynamique du répertoire T présente un fort avantage sélectif, car il permet d'accorder diversité du soi et diversité des répertoires générés au sein de chaque organisme individuel.

La commutation CD4+/CD8+

Le développement d'une cellule DP en cellule SP CD4+ ou SP CD8+ dépend de la classe de CMH, I ou II qui a permis sa sélection positive. Les thymocytes D,P positivement sélectionnés sur un CMH de classe I, perdent le co-récepteur CD4 et retiennent le co-récepteur CD8, s'engageant dans la lignée lymphocytaire T cytotoxique. Inversement, les lymphocytes T positivement sélectionnés sur un CMH-II deviennent des cellules Simple Positives (SP) CD4+ et mèneront aux lymphocytes T *helpers*. Au niveau du mécanisme, une première interaction du TCR avec une classe de CMH donnée inhiberait transitoirement le co-récepteur antagoniste et participerait à la sélection du TCR, puis une deuxième interaction orienterait la cellule vers la lignée correspondant au CMH reconnu lors du premier contact [Brugnera et al. 2000].

Effet de la commutation CD4+/CD8+ sur les répertoires

Des études ont identifié une expression différentielle de certaines chaînes TCRβ et TCRα à la surface des cellules CD4+ et CD8+ [Sim et al. 1998]. Pour rappel, les chaînes de TCRα et de TCRβ sont formées à partir des gènes V(D)J, le gène V codant pour les deux boucles CDR1 et CDR2, engagées dans la reconnaissance du CMH-peptide. L'utilisation des gènes V des loci TRA et TRB est différente entre les compartiments lymphocytaires CD4+ et CD8+. Pour plus de précision, les familles de gènes TRBV11 mais aussi TRADV2, TRADV3, TRADV11 sont plus utilisées dans le cas des cellules CD4+ thymiques, alors que les gènes variables BV4 et ADV8 sont majoritairement présents dans les cellules CD8+ [Pircher et al. 1992; Sim et al. 1998]. Les boucles CDR1 et CDR2 synthétisées à partir de ces familles de gènes V s'associent préférentiellement avec les CMH de type I et II, expliquant ce biais de répertoire [Sim, Zerva et al. 1996; Garcia et al. 1999].

Effet de la prolifération post-sélection thymique sur le répertoire

Les cellules SP nouvellement sélectionnées procèdent à une expansion clonale de 5 à 6 cycles de division [Ceredig 1990]. Cette multiplication cellulaire (dépendante de l'IL7 et de l'IL2 durant l'ontogénie précoce) ne semble pas être dépendante d'une reconnaissance de l'association CMH-peptide par le TCR, et cette expansion clonale n'entraînerait donc pas de biais du répertoire. Cette multiplication cellulaire correspondrait à un mécanisme homéostatique, modulant le nombre de lymphocytes issus du thymus, en fonction des besoins de la périphérie.

Références du Chapitre I

Abarrategui I, Krangel MS. Regulation of T cell receptor-alpha gene recombination by transcription. Nat Immunol. 2006. 7:1109-15.

Aidinis V, Bonaldi T, et al. The RAG1 homeodomain recruits HMG1 and HMG2 to facilitate recombination signal sequence binding and to enhance the intrinsic DNA-bending activity of RAG1-RAG2. Mol Cell Biol. 1999. 19: 6532-42.

Aifantis I, Buer J, et al. Essential role of the pre-T cell receptor in allelic exclusion of the T cell receptor beta locus. Immunity. 1997. 7: 601-7.

Akamatsu Y, Tsurushita N, et al. Essential residues in V(D)J recombination signals. J Immunol. 1994.153: 4520-9.

Akashi K, Traver D, et al. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature. 2000. 404: 193-7.

Alam SM and Gascoigne NR. Posttranslational regulation of TCR Valpha allelic exclusion during T cell differentiation. J Immunol. 1998. 160: 3883-90.

Anderson G, Hare KJ, et al. Positive selection of thymocytes: the long and winding road." Immunol Today. 1999. 20: 463-8.

Artavanis-Tsakonas S, Rand MD, Lake RJ. Notch signaling: cell fate control and signal integration in development. Science. 1999. 284:770-6.

Balciunaite G, Ceredig R, Rolink AG. The earliest subpopulation of mouse thymocytes contains potent T, significant macrophage, and natural killer cell but no B-lymphocyte potential. Blood. 2005. 105:1930-6.

Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. Cell. 2002. Suppl:S45-55.

Baumann M, Mamais A, McBlane F, Xiao H, Boyes J. Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. EMBO J. 2003. 22:5197–5207.

Becker PB, Horz W. ATP-dependent nucleosome remodeling. Annu Rev Biochem. 2002.71:247–273.

Benoist, C. and D. Mathis. Positive selection of T cells: fastidious or promiscuous? Curr Opin Immunol. 1997. 9: 245-9.

Bertocci B, De Smet A, Weill JC, Reynaud CA. Nonoverlapping functions of DNA polymerases mu, lambda, and terminal deoxynucleotidyltransferase during immunoglobulin V(D)J recombination in vivo. Immunity. 2006. 25:31-41.

Besseyrias V, Fiorini E, Strobl LJ, Zimber-Strobl U, Dumortier A, Koch U, Arcangeli ML, Ezine S, Macdonald HR, Radtke F. Hierarchy of Notch-Delta interactions promoting T cell lineage commitment and maturation. J Exp Med. 2007. 204:331-43.

Bhandoola A, von Boehmer H, Petrie HT, Zúñiga-Pflücker JC. Commitment and developmental potential of extrathymic and intrathymic T cell precursors: plenty to choose from. Immunity. 2007. 26:678-89.

Bhandoola A, Cibotti R, et al. Positive selection as a developmental progression initiated by alpha beta TCR signals that fix TCR specificity prior to lineage commitment. Immunity. 1999 10: 301-11.

Blom B, Spits H. Development of human lymphoid cells. Annu Rev Immunol. 2006. 24:287-320.

Bolland DJ, Wood AL, Johnston CM, Bunting SF, Morgan G, Chakalova L, Fraser PJ, Corcoran AE. Antisense intergenic transcription in V(D)J recombination. Nat Immunol. 2004. 5:630-7.

Bray SJ. Notch signalling: a simple pathway becomes complex. Nat Rev Mol Cell Biol. 2006. 7:678-89.

Buck D, Malivert L, de Chasseval R, Barraud A, Fondanèche MC, Sanal O, Plebani A, Stéphan JL, Hufnagel M, le Deist F, Fischer A, Durandy A, de Villartay JP, Revy P. Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. Cell. 2006. 124:287-99.

Buck D, Moshous D, de Chasseval R, Ma Y, le Deist F, Cavazzana-Calvo M, Fischer A, Casanova JL, Lieber MR, de Villartay JP. Severe combined immunodeficiency and microcephaly in siblings with hypomorphic mutations in DNA ligase IV. Eur J Immunol. 2006. 36:224-35.

Buer J, Aifantis I, et al. Role of different T cell receptors in the development of pre-T cells. J Exp Med. 1997. 185:1541-7.

Brugnera E, Bhandoola A, et al. Coreceptor reversal in the thymus: signaled CD4+8+ thymocytes initially terminate CD8 transcription even when differentiating into CD8+ T cells. Immunity. 2000. 13:59-71.

Capone M, Hockett RD Jr, et al. Kinetics of T cell receptor beta, gamma, and delta rearrangements during adult thymic development: T cell receptor rearrangements are present in CD44(+)CD25(+) Pro-T thymocytes. Proc Natl Acad Sci USA. 1998. 95: 12522-7.

Ceredig R, Rolink T. A positive look at double-negative thymocytes. Nat Rev Immunol. 2002. 2:888-97.

Ceredig R. Intrathymic proliferation of perinatal mouse alpha beta and gamma delta T cell receptor-expressing mature T cells. Int Immunol. 1990. 2:859-67.

Chakraborty T, Chowdhury D, Keyes A, Jani A, Subrahmanyam R, Ivanova I, Sen R. Repeat organization and epigenetic regulation of the DH-Cmu domain of the immunoglobulin heavy-chain gene locus. Mol Cell. 2007. 27:842-50.

Ciubotaru M, Kriatchko AN, Swanson PC, Bright FV, Schatz DG. Fluorescence resonance energy transfer analysis of recombination signal sequence configuration in the RAG1/2 synaptic complex. Mol Cell Biol. 2007. 27:4745-58.

Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM. Accessibility control of V(D)J recombination. Adv Immunol. 2006. 91:45-109.

Coleclough C. Chance, necessity and antibody gene dynamics. Nature. 1983. 303:23-6.

Cooper MD, Chen CL, Bucy RP, Thompson CB. Avian T cell ontogeny. Adv Immunol 1991. 50:87-117.

Corcoran AE, Riddell A, Krooshoop D, Venkitaraman AR. Impaired immunoglobulin gene rearrangement in mice lacking the IL-7 receptor. Nature. 1998. 391:904-7.

Corthay A, Nandakumar KS, Holmdahl R. Evaluation of the percentage of peripheral T cells with two different T cell receptor alpha-chains and of their potential role in autoimmunity. J Autoimmun. 2001. 16:423-9.

Davodeau, F., M. Difilippantonio, et al. The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. Embo J. 2001. 20:4717-29.

De P, Rodgers KK. Putting the pieces together: identification and characterization of structural domains in the V(D)J recombination protein RAG1. Immunol Rev. 2004. 200:70-82.

Dik WA, Pike-Overzet K, Weerkamp F, de Ridder D, de Haas EF, Baert MR, van der Spek P, Koster EE, Reinders MJ, van Dongen JJ, Langerak AW, Staal FJ. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med. 2005. 201:1715-23.

Dudley EC, Petrie HT, et al. T cell receptor beta chain gene rearrangement and selection during thymocyte development in adult mice. Immunity. 1994. 1:83-93.

Dujka ME, Puebla-Osorio N, Tavana O, Sang M, Zhu C. ATM and p53 are essential in the cell-cycle containment of DNA breaks during V(D)J recombination in vivo. Oncogene. 2010. 29:957-65.

Eastman QM, Leu TM, Schatz DG. Initiation of V(D)J recombination in vitro obeying the 12/23 rule. Nature. 1996. 380:85-8.

Elliott JI and Altmann DM. Dual T Cell Receptor alpha Chain T Cells In Autoimmunity. J. Exp. Med.1995. 182 :953-960

Faro-Trindade I, Cook PR. Transcription factories: structures conserved during differentiation and evolution. Biochem Soc Trans. 2006. 34:1133-7.

Feeney AJ, Tang A, Ogwaro KM. B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. Immunol Rev. 2000. 175:59-69.

Fugmann, SD, Lee AI, et al. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. Annu Rev Immunol. 2000. 18:495-527.

Garcia KC, Teyton L, et al. Structural basis of T cell recognition. Annu Rev Immunol. 1999. 17:369-97.

Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem. 2002. 71:101-32.

Georgescu C, Longabaugh WJ, Scripture-Adams DD, David-Fung ES, Yui MA, Zarnegar MA, Bolouri H, Rothenberg EV. A gene regulatory network armature for T lymphocyte specification. Proc Natl Acad Sci USA. 2008. 105:20100-5.

Gilfillan S, Dierich A, Lemeur M, Benoist C, Mathis D. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. Science. 1993. 261:1175-8.

Glusman G, Rowen L, et al. Comparative genomics of the human and mouse T cell receptor loci. Immunity. 2001. 15: 337-49.

Godfrey DI, Kennedy J, et al. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. J Immunol. 1993. 150: 4244-52.

Golding A, Chandler S, Ballestar E, Wolffe AP, Schlissel MS. Nucleosome structure completely inhibits in vitro cleavage by the V(D)J recombinase. EMBO J. 1999.18:3712–3723.

Görisch SM, Wachsmuth M, Tóth KF, Lichter P, Rippe K. Histone acetylation increases chromatin accessibility. J Cell Sci. 2005. 118:5825-34.

Hare KJ, Wilkinson RW, et al. Identification of a developmentally regulated phase of postselection expansion driven by thymic epithelium. J Immunol. 1998. 160:3666-72.

Haren L, Ton-Hoang B, Chandler M. Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol. 1999. 53:245-81.

Hayday AC. γδ cells: a right time and a right place for a conserved third way of protection. Annu Rev Immunol. 2000. 18:975-1026

Hein WR, Mackay CR: Prominence of gamma delta T cells in the ruminant immune system. Immunol Today. 1991. 12:30-34.

Herzig C, Blumerman S, Lefranc M-P, Baldwin C: Bovine T cell receptor gamma variable and constant genes: combinatorial usage by circulating γδ T cells. Immunogenetics 2006. 58:138-151.

Hong L, Schroth GP, et al. Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA. J Biol Chem. 1993. 268:305-14.

Huang J, Durum SK, Muegge K. Cutting edge: histone acetylation and recombination at the TCR gamma locus follows IL-7 induction. J. Immunol. 2001. 167:6073–6077.

Hughes MM, Yassai M, et al. T cell receptor CDR3 loop length repertoire is determined primarily by features of the V(D)J recombination reaction. Eur J Immunol. 2003. 33:1568-75.

Johnson K, Pflugh DL, Yu D, Hesslein DG, Lin KI, Bothwell AL, Thomas-Tikhonenko A, Schatz DG, Calame K. B cell-specific loss of histone 3 lysine 9 methylation in the V(H) locus depends on Pax5. Nat Immunol. 2004. 5:853-61.

Jones JM, Gellert M. Ordered assembly of the V(D)J synaptic complex ensures accurate recombination. EMBO J. 2002. 21:4162-71.

Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell. 2010. 141:419-31.

Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet. 1998. 19:187-91.

Kang YH, Son CY, Lee CH, Ryu CJ. Aberrant V(D)J cleavages in T cell receptor beta enhancer- and p53-deficient lymphoma cells. Oncol Rep. 2010. 23:1463-8.

Karsunky H, Inlay MA, Serwold T, Bhattacharya D, Weissman IL. Flk2+ common lymphoid progenitors possess equivalent differentiation potential for the B and T lineages. Blood. 2008. 111:5562-70.

Kaye J, Hsu ML, et al. Selective development of CD4+ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. Nature. 1989. 341: 746-9.

Kim DR, Dai Y, Mundy CL, Yang W, Oettinger MA. Mutations of acidic residues in RAG1 define the active site of the V(D)J recombinase. Genes Dev. 1999. 13:3070-80.

Klein L and Kyewski B. Self-antigen presentation by thymic stromal cells: a subtle division of labor. Curr Opin Immunol. 2000. 12:179-86.

Komori T, Okada A, Stewart V, Alt FW. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. Science. 1993. 261:1171-5.

Kondo M, Weissman IL, Akashi K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell. 1997. 91:661-72.

Kosak ST, Skok JA, Medina KL, Riblet R, Le Beau MM, Fisher AG, Singh H. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science. 2002. 296:158-62.

Krangel MS. Gene segment selection in V(D)J recombination: accessibility and beyond. Nat Immunol. 2003. 4:624-30.

Kwon J, Morshead KB, Guyon JR, Kingston RE, Oettinger MA. Histone acetylation and hSWI/SNF remodeling act in concert to stimulate V(D)J cleavage of nucleosomal DNA. Mol. Cell. 2000. 6:1037-1048.

Kwon J, Imbalzano AN, Matthews A, Oettinger MA. Accessibility of nucleosomal DNA to V(D)J cleavage is modulated by RSS positioning and HMG1. Mol. Cell. 1998. 2:829–839.

Lai AY, Kondo M. Identification of a bone marrow precursor of the earliest thymocytes in adult mouse. Proc Natl Acad Sci USA. 2007. 104:6311-6.

Laufer TM, Glimcher LH, et al. Using thymus anatomy to dissect T cell repertoire selection. Semin Immunol. 1999. 11: 65-70.

Lennon GG, Perry RP. C mu-containing transcripts initiate heterogeneously within the IgH enhancer region and contain a novel 5'-nontranslatable exon. Nature. 1985. 318:475-8.

Liu Y, Zhang L, Desiderio S. Temporal and spatial regulation of V(D)J recombination: interactions of extrinsic factors with the RAG complex. Adv Exp Med Biol. 2009. 650:157-65.

Liu Y, Subrahmanyam R, Chakraborty T, Sen R, Desiderio S. A plant homeodomain in RAG-2 that binds Hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. Immunity. 2007. 27:561-71.

Ma Y, Pannicke U, Schwarz K, Lieber MR. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell. 2002. 108:781-94.

Maillard I, Fang T, Pear WS. Regulation of lymphoid development, differentiation, and function by the Notch pathway. Annu Rev Immunol. 2005. 23:945-74.

Malissen M, Trucy J, et al. Regulation of TCR alpha and beta gene allelic exclusion during T-cell development. Immunol Today. 1992. 13:315-22.

Mancini S, Candeias SM, et al. TCR alpha-chain repertoire in pTalpha-deficient mice is diverse and developmentally regulated: implications for pre-TCR functions and TCRA gene rearrangement. J Immunol. 1999. 163:6053-9.

Matthews AG, Kuo AJ, Ramón-Maiques S, Han S, Champagne KS, Ivanov D, Gallardo M, Carney D, Cheung P, Ciccone DN, Walter KL, Utz PJ, Shi Y, Kutateladze TG, Yang W, Gozani O, Oettinger MA.

RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature. 2007. 450:1106-10.

Mattick JS. The functional genomics of noncoding RNA.Science. 2005. 309:1527-8.

McBlane F, Boyes J. Stimulation of V(D)J recombination by histone acetylation. Curr. Biol. 2000. 10:483–486.

McBlane JF, van Gent DC, Ramsden DA, Romeo C, Cuomo CA, Gellert M, Oettinger MA. Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. Cell. 1995. 83:387-95.

McMurry MT, Krangel MS. A role for histone acetylation in the developmental regulation of VDJ recombination. Science. 2000. 287:495-498.

Meek K, Dang V, Lees-Miller SP. DNA-PK: the means to justify the ends? Adv. Immunol. 2008. 99:33–58.

Monroe RJ, Chen F, et al. RAG2 is regulated differentially in B and T cells by elements 5' of the promoter. Proc Natl Acad Sci USA. 1999. 96:12713-8.

Morshead KB, Ciccone DN, Taverna SD, Allis CD, Oettinger MA. Antigen receptor loci poised for V(D)J rearrangement are broadly associated with BRG1 and flanked by peaks of histone H3 dimethylated at lysine 4. Proc Natl Acad Sci USA. 2003. 100:11577-82.

Mundy CL, Patenge N, Matthews AG, Oettinger MA. Assembly of the RAG1/RAG2 synaptic complex. Mol Cell Biol. 2002. 22 :69-77.

Narlikar GJ, Fan HY, Kingston RE. Cooperation between complexes that regulate chromatin structure and transcription. Cell. 2002. 108:475-87.

Naumov YN, Naumova EN, et al. A fractal clonotype distribution in the CD8+ memory T cell repertoire could optimize potential for immune responses. J Immunol. 2003. 170:3994-4001.

Nightingale KP, Baumann M, Eberharter A, Mamais A, Becker PB, Boyes J. Acetylation increases access of remodelling complexes to their nucleosome targets to enhance initiation of V(D)J recombination. Nucleic Acids Res. 2007. 35:6311–6321.

Oettinger MA, Schatz DG, et al. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science. 1990. 248:1517-23.

Osipovich O, Milley R, Meade A, Tachibana M, Shinkai Y, Krangel MS, Oltz EM. Targeted inhibition of V(D)J recombination by a histone methyltransferase. Nat Immunol. 2004. 5:309-16.

Pannetier C, Cochet M, et al. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. Proc Natl Acad Sci USA. 1993. 90:4319-23.

Pannicke U, Ma Y, Hopfner KP, Niewolik D, Lieber MR, Schwarz K. Functional and biochemical dissection of the structure-specific nuclease ARTEMIS. EMBO J. 2004. 23:1987-97.

Pearse M, Wu L, et al. A murine early thymocyte developmental sequence is marked by transient expression of the interleukin 2 receptor. Proc Natl Acad Sci USA. 1989. 86:1614-8.

Perry SS, Welner RS, Kouro T, Kincade PW, Sun XH. Primitive lymphoid progenitors in bone marrow with T lineage reconstituting potential. J Immunol. 2006. 177:2880-7.

Pircher H, Rebai N, et al. Preferential positive selection of V alpha 2+ CD8+ T cells in mouse strains expressing both H-2k and T cell receptor V alpha a haplotypes: determination with a V alpha 2-specific monoclonal antibody. Eur J Immunol. 1992. 22:399-404.

Porritt HE, Rumfelt LL, Tabrizifard S, Schmitt TM, Zúñiga-Pflücker JC, Petrie HT. Heterogeneity among DN1 prothymocytes reveals multiple progenitors with different capacities to generate T cell and non-T cell lineages. Immunity. 2004. 20:735-45.

Ragoczy T, Bender MA, Telling A, Byron R, Groudine M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. Genes Dev. 2006. 20:1447-57.

Ramsden DA, Baetz K, Wu GE. Conservation of sequence in recombination signal sequence spacers. Nucleic Acids Res. 1994. 22:1785-96.

Read AP, Strachan T. Chapter 18: Cancer Genetics". Human molecular genetics 2. New York: Wiley. 1999. ISBN 0-471-33061-2.

Rock EP, Sibbald PR, et al. CDR3 length in antigen-specific immune receptors. J Exp Med. 1994. 179:323-8.

Roldán E, Fuxa M, Chong W, Martinez D, Novatchkova M, Busslinger M, Skok JA. Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. Nat Immunol. 2005. 6:31-41.

Roth DB, Roth SY. Unequal access: regulating V(D)J recombination through chromatin remodeling. Cell. 2000. 103:699–702.

Roth DB, Menetski JP, Nakajima PB, Bosma MJ, Gellert M. V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. Cell. 1992. 70:983-91.

Rothenberg EV, Taghon T. Molecular genetics of T cell development. Annu Rev Immunol. 2005. 23:601-49.

Sayegh CE, Jhunjhunwala S, Riblet R, Murre C. Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. Genes Dev. 2005. 19:322-7.

Schatz DG, Oettinger MA, Baltimore D. The V(D)J recombination activating gene, RAG-1. Cell. 1989. 59:1035-48.

Schatz DG, Baltimore D. Stable expression of immunoglobulin gene V(D)J recombinase activity by gene transfer into 3T3 fibroblasts. Cell. 1988. 53:107-15.

Schlissel M, Constantinescu A, Morrow T, Baxter M, Peng A. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. Genes Dev. 1993. 7:2520-32.

Schwarz BA, Bhandoola A. Circulating hematopoietic progenitors with T lineage potential. Nat Immunol. 2004. 5:953-60.

Shimazaki N, Tsai AG, Lieber MR.H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. Mol Cell. 2009. 34:535-44.

Shimonkevitz R, Kappler J, Marrack P, Grey H. Antigen recognition by H-2-restricted T cells. I. Cell-free antigen processing. J Exp Med. 1983. 158:303-16.

Shinkai Y, Rathbun G, et al. RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. Cell. 1992. 68:855-67.

Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, Peterson CL. Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science. 2006. 311:844-7.

Sim BC, Lo D, et al. Preferential expression of TCR V alpha regions in CD4/CD8 subsets: class discrimination or co-receptor recognition? Immunol Today. 1998. 19: 276-82.

Sim BC, Wung JL, et al. Polymorphism within a TCRAV family influences the repertoire through class I/II restriction. J Immunol. 1998. 160:1204-11.

Sim BC, Zerva L, et al. Control of MHC restriction by TCR Valpha CDR1 and CDR2. Science. 1996. 273:963-6.

Singh H, Medina KL, Pongubala JM. Contingent gene regulatory networks and B cell fate specification. Proc Natl Acad Sci USA. 2005. 102:4949-53.

Skok JA, Gisler R, Novatchkova M, Farmer D, de Laat W, Busslinger M. Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes. Nat Immunol. 2007. 8:378-87.

Sprent J, Lo D, et al. T cell selection in the thymus. Immunol Rev. 1988. 101:173-90.

Surh CD, Lee DS, et al. Thymic selection by a single MHC/peptide ligand produces a semidiverse repertoire of CD4+ T cells. Immunity. 1997. 7:209-19.

Swanson PC. The bounty of RAGs: recombination signal complexes and reaction outcomes. Immunol Rev. 2004. 200:90-114.

Swanson PC. Fine structure and activity of discrete RAG-HMG complexes on V(D)J recombination signals. Mol Cell Biol. 2002. 22:1340-51.

Swanson PC. The DDE motif in RAG-1 is contributed in trans to a single active site that catalyzes the nicking and transesterification steps of V(D)J recombination. Mol Cell Biol. 2001.21:449-58.

Thomas JO. HMG1 and 2: architectural DNA-binding proteins. Biochem Soc Trans. 2001. 29:395-401.

Thompson A, Timmers E, Schuurman RK, Hendriks RW. Immunoglobulin heavy chain germ-line JH-C mu transcription in human precursor B lymphocytes initiates in a unique region upstream of DQ52. Eur J Immunol. 1995. 25:257-61.

Till JE, McCulloch EA. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiation Research. 1961. 14:213-22.

Tonegawa S. Somatic generation of antibody diversity. Nature. 1983. 302:575-81.

Tourigny MR, Mazel S, et al. T cell receptor (TCR)-beta gene recombination: dissociation from cell cycle regulation and developmental progression during T cell ontogeny. J Exp Med. 1997. 185:1549-56.

Tourne S, Miyazaki T, et al. Selection of a broad repertoire of CD4+ T cells in H-2Ma0/0 mice. Immunity. 1997. 7:187-95.

van Gent DC, van der Burg M. Non-homologous end-joining, a sticky affair. Oncogene. 2007. 26:7731-40.

van Gent DC, Hiom K, Paull TT, Gellert M. Stimulation of V(D)J cleavage by high mobility group proteins. EMBO J. 1997. 16:2665-70.

Viola A, Lanzavecchia A. T cell activation determined by T cell receptor number and tunable thresholds. Science. 1996. 273:104-6.

Wada H, Masuda K, Satoh R, Kakugawa K, Ikawa T, Katsura Y, Kawamoto H. Adult T-cell progenitors retain myeloid potential. Nature. 2008. 452:768-72.

Williams O, Tanaka Y, et al. Inhibition of thymocyte negative selection by T cell receptor antagonist peptides. Eur J Immunol. 1996. 26:532-8.

Wilson A, Marechal C, et al. Biased V beta usage in immature thymocytes is independent of DJ beta proximity and pT alpha pairing. J Immunol. 2001. 166:51-7.

Wolffe AP, Guschin D. Review: chromatin structural features and targets that regulate transcription. J Struct Biol. 2000. 129:102-22.

Yancopoulos GD, Alt FW. Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell. 1985. 40:271-81.

Yaneva M, Kowalewski T, et al. Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies. Embo J. 1997. 16:5098-112.

Ye SK, Agata Y, Lee HC, Kurooka H, Kitamura T, Shimizu A, Honjo T, Ikuta K. The IL-7 receptor controls the accessibility of the TCRgamma locus by Stat5 and histone acetylation. Immunity. 2001. 15:813-23.

Yoshida T, Tsuboi A, et al. The DNA-bending protein, HMG1, is required for correct cleavage of 23 bp recombination signal sequences by recombination activating gene proteins in vitro. Int Immunol. 2000. 12:721-9.

Generalities on Biomodeling

Généralités à propos de la modélisation en Biologie.

La modélisation conceptuelle en biologie ou biomodélisation, est une approche mathématique qualitative ou quantitative dirigée vers la compréhension des systèmes biologiques. Dans un sens général, il s'agit d'une activité cognitive qui décrit formellement la réalité biologique. Par la suite, nous utilisons quelques considérations publiées par des chercheurs partageant les mêmes points de vue à propos de l'approche de biomodélisation [Demongeot *et al*, 2003].

Conceptual modeling in biology, or biomodeling, is a qualitative or quantitative mathematical approach aiming to understand biological systems; in a general meaning, it constitutes a cognitive activity describing formally the biological reality. In the following, we will use some considerations published by researchers sharing the same viewpoints about the biomodeling approach [Demongeot *et al*, 2003].

A first advantage of the biomodeling approach stands even before data acquisition. In fact, the model-driven acquisition is aimed to the optimization of experimental measures through the identification of some crucial variables or critical parameter by a model results analysis step. The crucial variables of a biological model are characterized by their high sensitivity to parameter changes: they display dramatic changes in their dynamical behavior for a small parametric perturbation (concerning their amplitudes and/or their periodicities if their dynamical regime is oscillatory). The model-driven experimental planning is designed for identifying the critical parameters, corresponding to the most important parameters of the biological process. Independently, crucial parameters can be used afterwards for the testing of hypotheses leading to validate or falsify a given conceptual model. Thus, the experimental planning and the conceptual modeling are closely related each together: the former helps to acquire data describing the biomedical reality, the latter provides a framework to interpret the observed data, organize the actions (e.g., therapy), and anticipate on potential evolutions (e.g., prognosis). The interest of these modeling study is manifest, especially when dealing for example with computer-assisted therapy, a specialty of our TIMC-IMAG laboratory [Taylor *et al*, 1996].

Relying on the conceptual mathematical model standing for the physiopathological mechanisms studied, come numerical simulations, which represent key issues in the biomodeling research. Numerical simulations stand for paradigms regarding direct and inverse problems, image analyses and synthesis or signal processing. By using different scenarios about the set of parameter values and initial conditions, the first goal of performing simulations consists in generating a realistic reproduction the observed behaviors. If the simulation reaches this goal, it means that all the mean parameters and more important behaviors of the studied process were included in the model. If simulation studies are mainly used to get more insights or understanding of systems, they also offer good advantages in planning experimentation, on processing real data, or when realization of experiments is impossible for ethical reasons or due to technical constraints. Numerical simulations have a direct link to the recent trend of virtual reality in biomedicine, where training and planning in preoperative situations and intraoperative guiding in computer-assisted interventions are central elements.

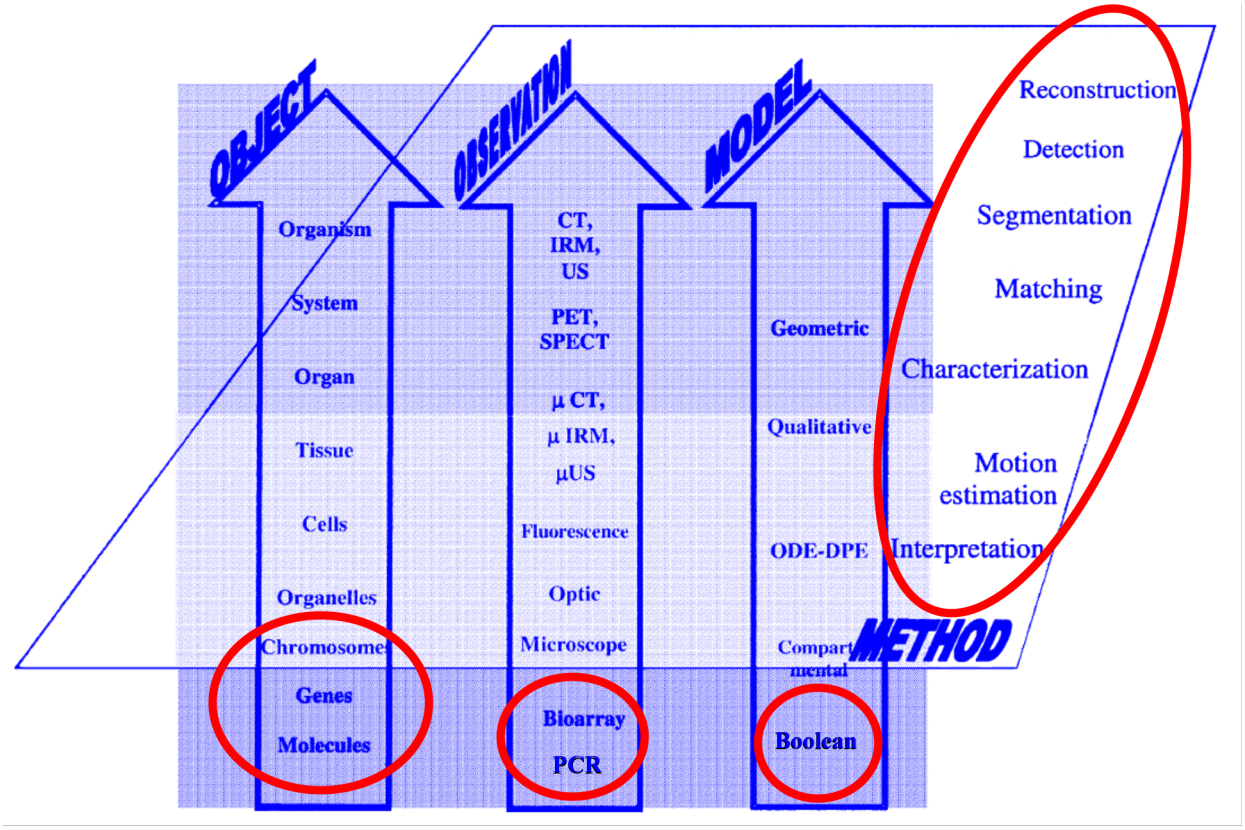


Figure 1. Object–observation–models. These basic dimensions, explored at different scale levels by means of appropriate acquisition modalities and models, can be submitted and coupled to generic components of information processing in order to gain in-depth understanding about normal and pathological features.

Modeling is indeed not limited to describe functions of elements constituting biological systems (Figure 1), but presents a major interest when used to find real solutions for everyday biomedical problems. The modeling approach is expected to provide the best resolution in its outputs when the latest knowledge in reasoning and decision processes are introduced explicitly in the model.

This resolution requirement applies to low level tasks (capture of information, control of processing algorithms,...), intermediate level actions (feature grouping, hypothesis generation,…), and high level analyses (diagnosis, corrective therapies, prognosis,...). When complex biological systems are concerned, explanations and causal relations are expected to come into sight together with explicit or numerical solutions to well-posed problems concerning empirical data interpretation.

Another ambitious goal in the biomodeling approach consists in identification or parameter estimation, which deals with both the system structure and the unmeasurable variables (estimating the internal parameters and variables of physiological interest). Identification makes possible prediction and control as well as the achievement of therapeutic agent administration tracking. However, theoretical problems remain still opened due to the high complexity level of biosystems. In fact, the observables are often sparse and obtained through a multi-level acquisition device network. The system identification often comes from several sources of merged data or does not belong to the general class of inverse problems. Identification still receives a limited attention from the signal processing community in spite of the two-approach proximity in terms of concepts or automatic controlled tool used. Inverse problems in biomedical engineering have been firstly addressed in image reconstruction [Herman *et al,* 1980], but now all medical fields are concerned by this strategy. In immunology for example, the inverse attitude leads to combine flow cytometry and multi-photon microscopy (for immuno-competent cells selection), PCR (for immune genome expression), familial pedigree collection, and information stored in dedicated databases (Figure 2).

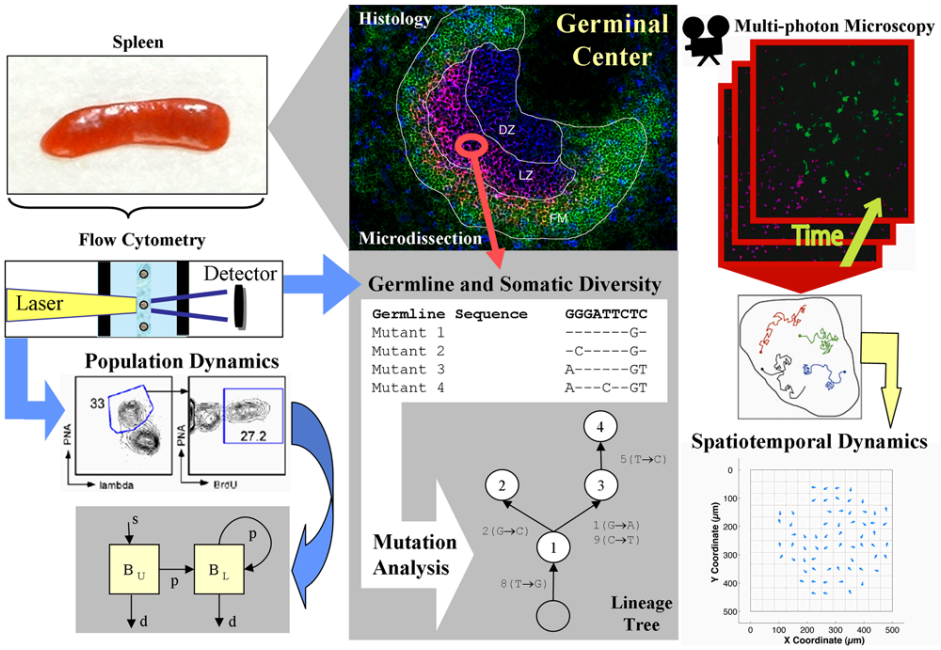


Figure 2. Multi-scale imaging in immunology. From [Kleinstein *et al*, 2008].

Systems Biology and Biocomplexity

The first challenge in the mathematical study of physiopathological processes comes from intrinsic system complexity and from the fact that numerous mechanisms are still not clearly established (as cellular controls involving large loop regulations in cell functions like proliferation, secretion, or motion). The biomodeling approach would be useful in increasing the system understanding by the capabilities models offer to deal with different scenarios of simulation (by changing either parameters values and/or initial conditions of variables). Comparing simulation outputs with empirical observations leads to predict behaviors and parameter values. In this way, models would be useful even in case of a complete understanding of the processes. Beyond, the studying of model parameter deviations from normal behavior (very complex behaviors may occur from structural changes in parameter set) may allow conjecturing the reverse process, meaning identifying ways to go back from abnormal to normal states. But there is surely a long way before being able to reliably reproduce the multitude of physiopathological behaviors in spite of the theory abundance.

Even basic mechanisms and interactions involve a large number of variables acting at different levels of organization (molecule, gene, protein, cell, organ, organism or individual, population) as well as strong reciprocal influences with the environmental conditions. The understanding of such a complexity, from very local entities to larger organizations motivated the emergence of the "integrative physiology" concept. Notably, if the integrative physiology were limited to the accumulation of knowledge advances coming from different fields of concern, it would present a reduced interest. For example sparse observations about electrical mechanisms (excitation, propagation), biomechanics (contraction / expansion / deformation estimated from motion in dynamic CT scanner or elastography form ultrasound), or fluid dynamics (blood flow imaging) will never be sufficient to get an integrated view of the heart, whatever the description level they convey. A real break through can be expected necessarily putting these different aspects in relation each together; hence models can play a major role for diagnosis, therapy or prognosis assessments. Indeed, the description of the relevant variable dynamics at each level and the modelization of the interactions between them can make emerge a more complete understanding of the processes. Finally, the essence of integrative physiology and more generally of systems biology consists in building a path from microscopic processes toward global properties, at macroscopic levels of organization.

Eventually, it can be said that models developed so far cannot afford any comprehensive

knowledge of the entire biological systems (a model quality called completeness). They are aimed to provide some insights on different aspects of the physiological or pathological processes.

Multilevel Models

New knowledge about cell or organ functions must be formalized through a high level of formal description (biochemical, mechanical, signaling processes used in cellular and tissular biology, genome expression, cell motion and tissue morphogenesis); thus, biological system studies often lead to multilevel integrated descriptions. This necessity to build more bridges between tissue, cell, molecular and higher organization levels relies on structural and functional information, which includes membrane features, channel kinetics, transporter characteristics, metabolic network topology as well as tissue properties (as architecture, elasticity, or fluid dynamics). Moreover, the close coupling between data, knowledge and models frequently allows revealing missing elements, ensuring a better coherence and pertinence in the explanation of observed physiological or pathological phenomena, which permits to increase further medical action efficiency.

Modeling Issues

Model complexity level

The plausibility of a model can be appreciated regarding the realistic features (architecture and functioning) it includes. A given model evaluation can be performed only with respect to another model: by the spectrum of generated simulation possibilities or by the amount of knowledge incorporated. The evaluation can be associated to the notion of completeness as well. A theoretical demonstration or proof that a model would be distinguishable from another is evidently advantageous, but other issues are equally decisive. In fact, minimal or parsimonious models present an interest: the falsifiability of a model is often due to its ability to account for the observed physiopathological behaviors in an economic way, and the "Occam's razor" principle stipulates that model entities (essentially variables and parameters) must not be multiplied beyond necessity ("entities must not be multiplied beyond necessity", from ["Occam's razor", 2003]). In such a situation of minimalist entities, it is often possible to conceive a crucial experiment able to falsify the model by contradicting one of its predictions, which leads to the conception of higher complexity status in the modeling approach. Thus, it is important to consider the inherent incompleteness of excessively complex models. Incorporation of too much complexity in a model leads to a high degree of imprecision, an uncertain model structure, and an unsure identification of parameter values. Consequently, the model should be difficult to validate. Nevertheless, the new "multi-agent" modeling approach, even being multi-level and imprecise in accounting for agent interactions, can be helpful if it shows the emergence of properties similar to those already described by lower description level models (e.g., from individual to population).

Data acquisition and model description level

It is important to notice that multiple factors may affect recorded data (sensor influence, noise, or non-stationarity) and can directly impact on model identifiability. In some cases, the type and modalities of measurements present strong restrictions. This is the case of compartmental models, which are applied to many functional imaging sources (MRI, SPECT, and PET) and have been specifically and extensively studied [Jacquez *et al*, 1996]. It is also the case of the bio-arrays, which offer complementary information on the "omic" regulatory networks, but whose interpretation highly depends on the number of observed variables: high

density DNA chips can contain too much genes to analyze lowering the model building reliability and therefore its predictions. In the contrary, when a too small number of variables are observed and incorporated in the model, simulations are unable to predict the very behaviors involving the missing variables. Important notions have been established regarding the *a priori* local and global identifiability for linear and nonlinear systems. For example, identification techniques for compartmental models may apply only on some specific cases, for instance: least squares or maximum likelihood when the model is uniquely identifiable; otherwise, in case of multiple identification, derivation of confidence interval bounds, parameter aggregation, and model reduction can help in estimating some of the parameters. Recent contributions have focused on differential algebra for nonlinear model identification [Audoly *et al*, 2001; Ljung *et al*, 1994].

Another critical issue is the description level (or scale), which highly influences the complexity of a model. This scale problem corresponds to different paradigms according to the research field (granularity, surface/deep, fine-to-coarse, local/global, detail level), and concerns both time and space scale selection. A practical rule to specify the convenient scale for a given model could rely on the tuning between three entities: i) the level of details or decomposition of the model; ii) the scale of observations; and iii) the available algorithmic processing capabilities (i.e, the quality of the information that it is possible to extract). Here also, the modeling purpose is of great importance: simulation will give more degrees of freedom (large scale, distributed systems) in studying different scenarios than a multiple identification, which would need parameter aggregation to allow their partial estimation.

Integration

The capability to integrate and merge exogenous data or information (clinical symptoms, electrical data, imaging examinations) is also of high relevance, but, with the exception of knowledge-based systems or statistical approaches, none of the theoretical models at our disposal is able to handle such requirements. The same comment emerge concerning heterogeneous model fusion (e.g., qualitative or symbolic with numerical models). More, in some models, the combinatorial space faced or the nonconvex functional minimized may lead to mathematical problems, independent from physiopathological concerns.

Model Classes

Thus, the definition of formal approaches comes beyond qualitative theories, with the use of the mathematical frameworks of dynamical systems. These systems are essentially of two types, i) discrete, that is with a finite number of states, observed in discrete time, and ii) continuous, with states in $\mathbb{R}^n$, observed in continuous time. Systems generate spatio-temporal patterns whose robustness to small modifications of the initial and boundary conditions, or to structural changes (parameter variations), presents a critical importance. In general, Ordinary Differential Equations (ODEs) are well suited to account for dynamical behaviors of individual entities or small size organizations, but they become intractable for large assemblies. Because of their intrinsic non-linearity, relevant equations for populations cannot be derived directly from the equations governing elementary components. In other words, models of individuals do not match models of assemblies; alternative options to ODE systems, like Boolean ones, will be examined and presented in the following.

In the Boolean approach, discrete variables represent the state of a gene (1 if expressed and 0 if silent) or the presence of an enzyme (1 if over a certain concentration threshold and 0 under). This formalization has been used extensively in regulatory network modeling [Kauffman, 1969; Thomas, 1973; Demongeot *et al*, 2002; Demongeot *et al*, 2003], and will be used for representing the presence or absence of accessible V or J genes in the successive windows of the TCRA rearrangement model presented Chapter 3.

Concerning differential continuous approaches, ordinary or partial differential equations (ODE or PDE) were commonly used by authors for studying the dynamics of the immunologic response. The continuous variables in these models are the sizes of the cellular immuno-competent populations and also the concentrations of the enzymes and metabolites, actors of the immune system, i.e. those causing the DNA rearrangement, the chromatin dynamics, the antibodies carrying and the transmembrane receptors implementation.

The selection of a model class (hidden Markov models, ODE, Boolean, or qualitative models) is among the early steps of a biomodeling approach. There is no objective rule making the decision, and no model class is able to claim genericity or superiority. The model choice is strongly conditioned by the nature of the observables and the available knowledge, the capability of a class to represent the processes, and the formal properties attached to the process (from identifiability to inversion problems), the time and finally, the space dimensions in which to operate. Any model, from external descriptions to physiologically based models, has a role to play for research or clinical applications as far as it is associated to a clear objective and a accurate validation process.

Of course, once modeling operation is performed, other issues emerge. Competition between approaches should be the basic principle, but unfortunately, the frequent absence of communication and cross-fertilization between disciplines most often prevents any comparison.

Most of the modeling issues can apply to all classes of models whatever their purposes. However, the ways modeling issues are answered depends on the nature of each model category. Modeling issues correspond mainly to important question marks and consequently to challenging problems to be solved and not to methodological answers [Carson *et al*, 2000]. Some are of concern only for mathematical models and of interest in the special case of closed-loop systems; a key property is the identifiability namely the possibility to estimate model parameters.

Model validation

The validation and/or falsification of models constitute a further complicated tasks in physiopathological systems modeling: it consists in estimating whether or not a model fits its purpose. The plausibility and quality properties of a model are difficult to evaluate. Establishing the validity range, either for simple or complex models, can be performed concerning cases, learning sets, experimental data matching, expert confrontations, or exception studies, but an objective-widely-accepted measure of validity is difficult to set. The validity assessment can rely on i) the model theoretical consistency, ii) its robustness to change or noise, iii) the model result deviations from experimental data, and iv) the testability of model results using protocols of practical tests. Moreover, the validity assessment should be appreciated through the model utility (assistance to a decision, capacity of prediction or conjecture formulation).

Finally, no permanent ground truth can be established directly from the experimental observation of living systems and diseases. A progressive scheme of knowledge validation may emerge from the use of fully simulated data, hybrid data (a mixture of synthetic data and real data) and real data with expert annotations. The ground truth is therefore partially controlled but its realism can constantly be questioned. Symmetrically, model-based simulations can be performed; statistical features on generated data can be extracted and compared to their real data equivalent.

Biomedical Triangle

It is interesting to consider the advantages of a high interconnection among acquisition

modalities and algorithmic resources: the combinatorial explosion existing behind the objects to modelize can only be faced through well posed questions (Figure 3). This is not an easy task because it requires a deep knowledge of the entire aspects of a given biomedical problem. The triangle and the bidirectional arrows in Figure 3 exemplify a partial space into which relevant and operational views have to be developed. For instance, *in-silico* simulations of DNA, protein, cell or organ behaviors can complement studies performed *in vitro* and *in vivo*: going from a medical problem, through the data acquisition, and toward the mining and modeling knowledge.

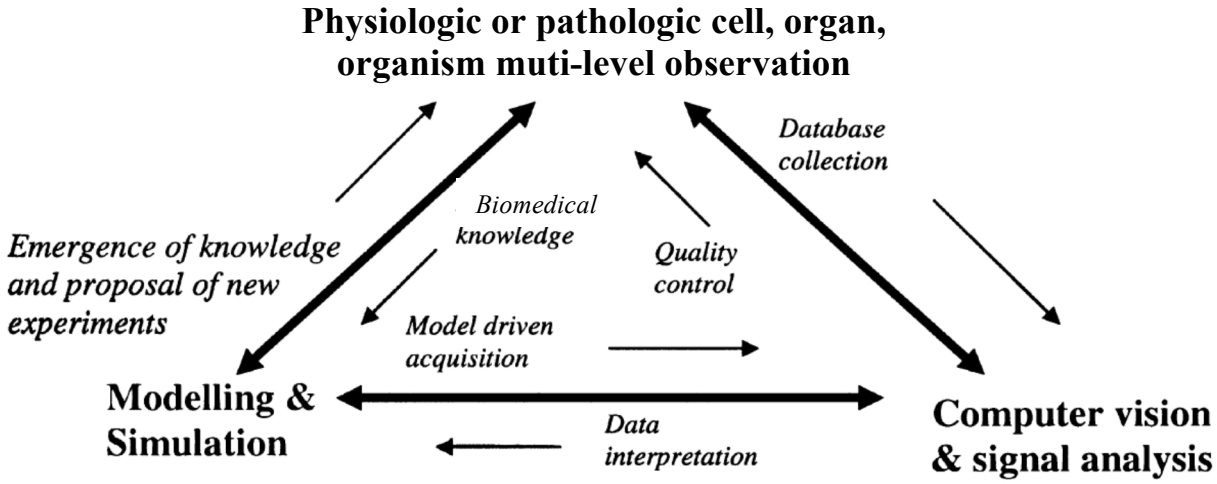


Figure 3. Interrelations depicted by the triangle point out the diverse facets corresponding to patient diagnosis and therapy, from medical knowledge, quality control up to database collection.

The need for a language able to describe the many different entities like cell functions, intra- or intercellular interactions, localization (growth and motion), signaling, cell physiology mechanisms, systems biology and cell anatomy is a major concern, and can be done in a common ontology framework (like Gene Ontology GO^{TM}). Such a universal language will allow the implementation of relevant algorithms, able to simulate functions, interactions, localization, or signaling with the use of hybrid models (e.g., Boolean systems for the genomic interactions, fixing the values of parameters of partial differential systems, whose attractors represent the final phenotypes to be explained).

All the modeling methods proposed in the following section deal with the richness of the immune repertoire and with the dynamics of the immune response; these modeling approaches are generic enough to deal with the several concepts and data available on mouse.

Classical models in immunology

Qualitative approach: the idiotypic network theory

In a famous paper, N.K. Jerne presented 35 years ago his idiotypic network theory about the principle of monoclonal antibody production: specific antibodies are supposed to be generated by one type of immune cells resulting from clonal expansion of a unique parent cell [Jerne, 1974]. This theory was the starting point of a large set of theories, claiming their capacities to explain the immunological tolerance, which is the ability of an individual to ignore its "self" proteins, while reacting against "non-self" ones. This breakage leads to the immune system's elaborating effective and specific immune responses against non-self determinants. The exact genesis of immunological tolerance is still elusive, but several theories have been proposed:

- Clonal Deletion theory, proposed by F.M. Burnet [Burnet, 1957], according to which self-reactive lymphoid cells are destroyed during the immune system development of an individual.
- Clonal Anergy theory, proposed by G. Nossal [Pike *et al*, 1982], in which self-reactive T- or B-cells become inactivated in the normal individual and cannot expand during an immune response.
- Idiotype Network theory, proposed by N.K. Jerne, wherein a network of antibodies able to neutralizing self-reactive antibodies is naturally present within the organism.
- The "Suppressor population" or "Regulatory T cells" theories [Lu *et al*, 2006], wherein regulatory T-lymphocytes (commonly $CD4^+FoxP3^+$ cells, among others) function to prevent, downregulate, or limit autoaggressive immune system responses.

More precisely, the idiotypic network theory corresponds to a qualitative theory in which the notion of network is central: autoantibodies are thought to recognize antibodies that recognize antibodies. According to this therory, an antigen generates antibodies whose serologically unique structure (idiotypic), results in the production of anti-idiotypic antibodies. Calling Ab_1 the original antibody and Ab_2 the anti-idiotypic antibody, the Ab_2 antibodies are believed to recognize the antigen-binding site of Ab_1, and therefore to share motif or structural similarities with the original antigen. The cascade then perpetuates, with the generation of anti-

anti-idiotypic antibodies (Ab_3) that recognize Ab_2, and so on [Shoenfeld *et al*, 2004]. The resulting end being supposedly the production of a chain of auto-antibody recognizing each other, which might modulate the immune system activity through its stimulation or suppression.

An elegant additional theory developed by D. Thomas [Padiolleau-Lefevre *et al*, 2003] proposes this cascade may state for the origin of the present functional proteins, which would have kept only essential functional sites through evolution, loosing supplementary peptidic chains being contingent for its function: therefore, the evolved functional protein are of the Ab_n form, with n even.

Dynamical models of the immunologic response (normal, paralysed, hyper reactive)

Independently of the discrete or continuous character of their formalizations, some classical models used in immunology will be now presented, illustrating the interest of a mathematical approach for testing some mechanistic scenarios concerning the anergy, paralysis or hyper-reactivity of the immune system, either against exogeneous infectious agents or against endogeneous proteins.

Perelson's modeling

In the study of immune cell dynamics, the selection that occurs during lymphocyte maturation may results in extensive cellular proliferation or death. Accurate rate measurements for these processes can help determining their relative contribution to the preferential expansion of higher-affinity B cell mutants. Dividing cells can be labeled using bromodeoxyuridine (BrdU), a thymidine analog that gets incorporated into DNA during S phase. The fraction of labeled cells is tracked during BrdU administration and following withdrawal using flow cytometry. To interpret the resulting data, Bonhoeffer *et al.* [Bonhoeffer *et al*, 2000] proposed a simple model that assumes a single B cell either proliferates in a clonal population according to a p rate, or undergoes apoptosis according to a d rate. In order to model BrdU labeling, this population is split into unlabeled (BU) and labeled (BL) subsets, whose concentrations are represented by x and y respectively. The dynamics of these two subpopulations is represented by the following linear equations:

$$dx/dt = s - (p+d)x, \qquad dy/dt = 2px - \mathrm{d}y,$$

where an unlabeled source of cells (s) and a 100% labeling efficiency were assumed. In [Kleinstein *et al*, 2008] is reported how Perelson modeled more than this simple proliferation phase, and in [Perelson *et al*, 1997], A.S. Perelson and G. Weisbuch formalized the cross-linking

of bivalent receptors by a bivalent antigen using a non-linear ODE; they analyzed a situation including a fixed number of B cells, each cell carrying a constant number of bivalent receptors S0 on its surface. Authors assumed the two sites on each receptor were identical and thus characterized through the same forward/reverse kinetic constants, and the same parameter values were applied whatever receptor is free or in an aggregate (equivalent-site hypothesis). Some variables were defined: $S(t)$ the concentration of free sites at time t, C the concentration of free bivalent antigen, C_1 the concentration of antigen bound at one site, and C_2 the concentration of ligand bound at both sites (C_2 corresponding to the concentration of cross links). To describe the kinetics of cross-linking, k_f and k_r were defined as kinetic constants describing the binding and release of one site on the antigen to and from a receptor site, more k_x and k_{-x} corresponded to kinetic constants describing the binding and unbinding of the second site on an antigen already bound at one site. With the corresponding equilibrium constants,
$K = k_f/k_r$ and $K_x = k_x/k_{-x}$, and by the law of mass action authors wrote:

$$C + S \underset{k_r}{\overset{2\,k_f}{\rightleftarrows}} C_1 \qquad\qquad C_1 + S \underset{2k_{-x}}{\overset{k_x}{\rightleftarrows}} C_2$$

or $$dC_1/dt = 2k_fCS - k_rC_1 - k_xC_1S + 2k_{-x}C_2, \qquad dC_2/dt = k_xC_1S - 2k_{-x}C_2$$

The factors of 2 arise because either site on the ligand can bind to a receptor site, and either of the two bound sites on C_2 can dissociate. By conservation of receptor sites,

$$S_0 = S(t) + C_1(t) + 2C_2(t),$$

where the fact that C1 occupied one site and C2 occupied two sites was used. In what followed, excess in ligand and constant quality of C(t) (denoted C for simplicity) were assumed, leading to the equilibrium solution:

$$C_1 = 2KCS,\ C_2 = K_xC_1S/2 = KK_xCS^2, \text{ hence } S = S_021\text{-}\beta)(\text{-}1+(1+\delta)^{1/2}/2\delta$$

This simple example was meant to illustrate how a bell-shaped cross-linking function can arise. An extensive modeling effort were then pursued, with the development of both dynamic and equilibrium models for cross-linking receptors considering bivalent, trivalent, and multivalent ligands, interacting with monovalent, bivalent, and multivalent receptors.

Segel's modeling

In [Bergmann *et al*, 2001], L.A. Segel addressed the crucial problem of the naive T cell fate decision between Th1 and Th2 cells, by describing the production rate equations for T helper populations in the following simple form:

Change Rate of antigen-specific Th1/Th2 cell population=differentiation+proliferation- death

T helper populations were polarized between Th1 and Th2 by an asymmetric cross-suppression. Whereas the Th1 cytokine IFN- inhibited proliferation of Th2 cells, Th2 secreted IL-10 suppressed cytokine production of Th1 cytokines, which indirectly down-regulated proliferation and differentiation of Th1, owing to a lack of growth factors and stimulators of differentiation (Figure 4).

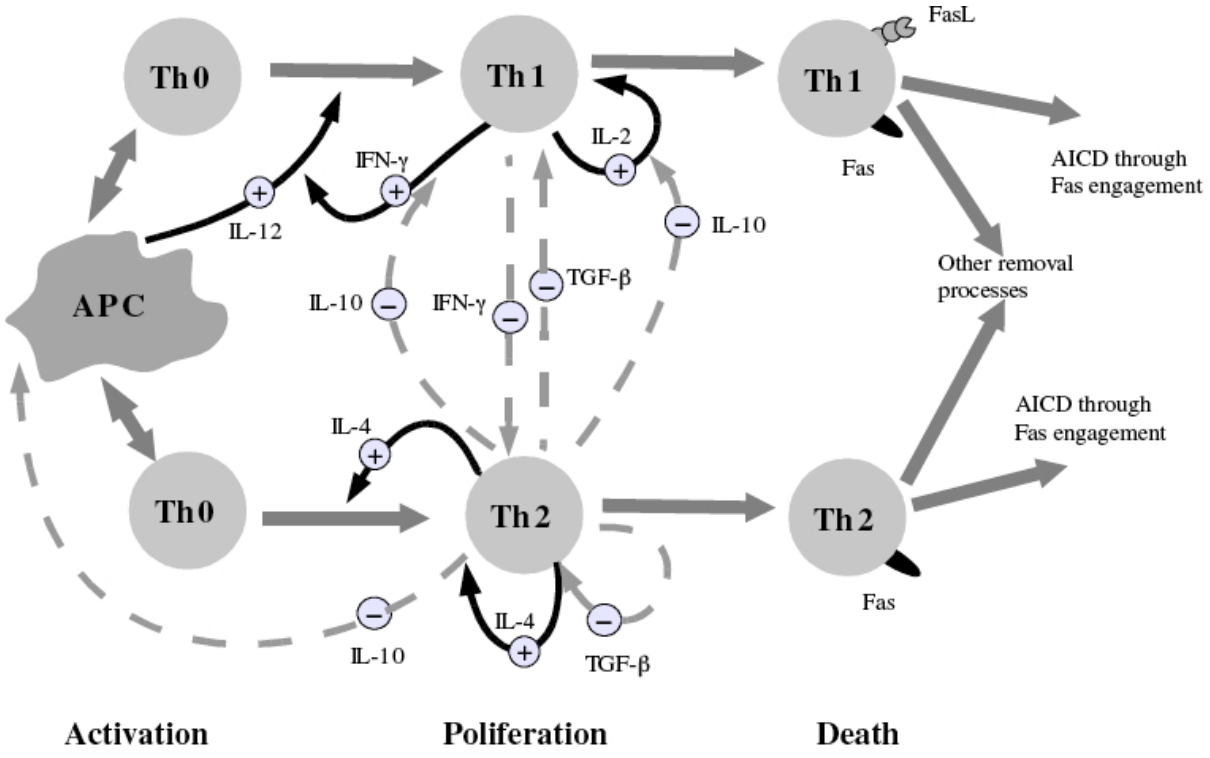


Figure 4. Schematic representation of the interactions governing the T helper system (AICD D antigen-induced cell death). After [Bergmann *et al*, 2001].

During all represented stages, T helper subsets are regulated by cytokines, produced by the corresponding or competing T helper type. Both T helper subsets exhibit positive feedback on their own differentiation from a naive Th0 state, implicating the involvement of the cytokines IL-4 for Th2 and a positive loop including antigen-presenting cells secreted IL-12 and IFN-γ for Th1. Authors assumed i) there was a sufficient pool of Th0 cells owing proliferation before they became committed to Th1 or Th2 cells (of total size respectively equal to T_1 and T_2) and ii) the majority of important cytokines involved would be classified as either type 1 or type 2, of concentrations S_1 and S_2, according to their predominantly production by Th1 or Th2. These assumptions were made in order to reduce the number of variables. The time scales of cytokine production, receptor binding, and cytokine decay being typically small compared to those of the

cell population dynamics, it lead to steady-state assumptions for S_1 and S_2, allowing to relate them directly to cell concentrations. Then the expressions for the cytokine signals took the form:

$$S_1 = \alpha_1 T_1/(1+k\alpha_2 T_2), \qquad S_2 = \alpha_2 T_2$$

where the α_i's are proportionality constants, and the equations of the differential system ruling the sizes of the Th1, Th2 and APC cells are given by:

$$dT_1/dt = \xi_1 P S_1/(1+kS_2) + \beta_1 T_1/(1+kS_2) - \delta_1 T_1^2 - \mu T_1,$$
$$dT_2/dt = \xi_2 P S_2/(1+kS_2) + \beta_2 T_2/(1+kS_1)(1+kS_2) - \delta_2 T_1^2 - \mu T_2,$$
$$dP/dt = \rho P - \omega_1 T_1 P - \omega_2 T_2 P,$$

where:
ξ_i is an activation strength, weighted for its Th*i*-inducing properties,
β_i is the efficiency of growth factors at maintaining activated cells in cycle
δ_i is the susceptibility of Th*i* cells to activation-induced cell death
ρ is the growth rate of pathogens
ω_i is the pathogen elimination efficiency of Th*i*-induced effectors.
μ is the death rate of Th1 and Th2 cells

The authors studied the conditions for existence and stability of the relevant steady states. From this analysis they reached conclusions of the following type:

- after pathogen elimination, the system ends up in a non-zero Th1-dominated or Th2-dominated steady state, which could be interpreted as memory states.
- if only one T helper type leads to pathogen destruction, a bistability may exist either with a successful Th1 response or a chronic Th2-dominated situation, or vice versa. This means that the very initial conditions fix on whether pathogen clearance or chronic disease is obtained.

Kaufmann's modeling

In [Kaufman *et al*, 1999], authors use the R. Thomas Boolean formalism [Thomas *et al*, 1973] in order to explain both positive (cell proliferation and cytokine production) and negative (anergy induction) signaling of T lymphocytes. The model relies on the autophosphorylative properties of the tyrosine kinases, which is associated with the T cell receptor. One of the basic assumptions is that the kinase activity of these receptor-associated enzymes remains above a background level after ligand removal and is responsible for a cellular unresponsiveness. Using the Boolean formalism, authors showed how the timing of the binding and intracellular signal-transduction events can affect the properties of receptor signaling and determine the type of cellular response, allowing making explicit the specification of conditions that lead to cellular activation or to anergy.

References of the Chapter II

Audoly S, Bellu G, D'Angio L, Saccomani MP, Cobelli C. Global identifiability of nonlinear models of biological systems. IEEE Trans. Biomed. Eng. 2001. 48 :55–65.

Bergmann C, van Hemmen JL, Segel LA. Th1 or Th2: How an Appropriate T Helper Response can be Made. Bulletin of Mathematical Biology. 2001. 63:405–430.

Bonhoeffer S, Mohri H, Ho D, Perelson AS. Quantification of cell turnover kinetics using 5-bromo-29 deoxyuridine. J. Immunol. 2000. 164 :5049–5054.

Burnet FM. A modification of Jerne's theory of antibody production using the concept of clonal selection. Australian Journal Science. 1957. 20:67–69.

Carson RE, Cobelli C. Modeling, Methodology for Physiology and Medicine. New York: Academic. 2000.

Demongeot J, Drouet E, Moreira A, Rechoum Y, Sené S. Micro-RNAs: viral genome and robustness of the genes expression in host. Phil. Trans. Royal Soc. 2009. A, 367 :4941-4965.

Demongeot J, Glade N, Moreira A, Vial L. RNA relics and origin of life. Int. J. Molecular Sciences. 2009. 10 :3420-3441.

Demongeot J, Bezy-Wendling J, Mattes J, Haigron P, Glade N, Coatrieux JL. Multiscale Modeling and Imaging: The Challenges of Biocomplexity. Proceedings of the IEEE Society. 2003. 91:1723-1737.

Demongeot J, Aracena J, Thuderoz F, Baum TP, Cohen O. Genetic regulation networks: circuits, regulons and attractors. C. R. Acad. Sci. Biologies. 2003. 326 :171-188.

Demongeot J, Thuderoz F, Baum TP, Berger F, Cohen O. Bio-array images processing and genetic networks modelling. C. R. Acad. Sci. Biologies. 2003. 326 :487-500.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modelling. ISBI 2002, IEEE EMB, M. Unser & Z.P. Liang, Eds. Piscataway: IEEE Press, 50-54 2002.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modelling, in : 5th IEEE EMBS International Summer School on Medical Imaging, J.L. Coatrieux *et al.,* Eds. Piscataway: IEEE Press, 15-23, 2002.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modelling, in : Modelling & simulation of biological processes in the context of genomics, P. Amar et al., Eds. Evry: Genopole, 87-94, 2002.

Fedeli M, Napolitano A, Man Wong MP, Marcais A, De Lalla C, Colucci F, Merkenschlager M, Dellabona P, Casorati G. Dicer-Dependent MicroRNA Pathway Controls Invariant NKT Cell Development. J. Immunology. 2009. 183 :2506 -2512.

Herman GT. Image Reconstruction From Projections. NewYork: Academic. 1980.

Jacquez J. Compartmental Analysis in Biology and Medicine. Ann Arbor, MI: BioMedware. 1996.

Jerne NK. Towards a network theory of the immune system. Ann. Immunol. 1974. 125:373-389.

Kaufman M, Andris F, Leo O. A logical analysis of T cell activation and anergy Proc. Natl. Acad. Sci. USA. 1999. 96 :3894-3899.

Kauffman S. Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets, J. Theoret. Biol. 1969. 22 :437–467.

Kleinstein SH. Getting Started in Computational Immunology. PLoS Comput. Biol. 2008. 4 : e1000128.

Ljung L, Glad ST. On global identifiability for arbitrary model parametrizations. Automatica. 1994. 30 :265–276.

Lodish HF, Zhou B, Liu G, Chen CZ. Micromanagement of the immune system by microRNAs. Nature Reviews Immunology. 2008. 8 :120-130.

Lu LF, Lind EF, Gondek DC, Bennett KA, Gleeson MW, Pino-Lagos K, Scott ZA, Coyle AJ, Reed JL, Van Snick J, Strom TB, Zheng XX, Noelle RJ. Mast cells are essential intermediaries in regulatory T-cell tolerance. Nature. 2006. 442:997-1002.

"Occam's razor". Merriam-Webster's Collegiate Dictionary (11th ed.). New York: Merriam-Webster. 2003. ISBN 0-87779-809-5.

Padiolleau-Lefevre S, Débat H, Thomas D, Friboulet A, Avalle B. In vivo evolution of a beta-lactamase-like activity throughout the idiotypic pathway. Biocat. Biotransf. 2003. 21 :79-85.

Perelson AS, Weisbuch G. Immunology for physicists. Rev. Mod. Phys. 1997. 69 :1219-1267.

Pike B, Boyd A, Nossal G. Clonal anergy: the universal anergic B lymphocyte. Proc Natl Acad Sci USA. 1982. 79:2013–2017.

Shoenfeld Y. The idiotypic network in autoimmunity: antibodies that bind antibodies that bind antibodies. Nature Medicine. 2004.10 :17-18.

Tassano E, Acquila M, Tavella E, Micalizzi C, Panarello C, Morerio C. MicroRNA-125b-1 and BLID upregulation resulting from a novel IGH translocation in childhood B-Cell precursor acute lymphoblastic leukemia. Genes Chromosomes Cancer. 2010. 49 :682-687.

Taylor R, Lavallée S, Burdea G, Mösges R. Computer-Integrated Surgery: Technology and Clinical Applications. Cambridge, MA: MIT Press. 1996.

Thomas R. Boolean formalization of genetic control circuits. J. Theoret. Biol. 1973. 42 :563-585.

Chapter III Dynamical Modeling of TRA/TRD Vα-Jα Rearrangements in Mice and Humans

Modélisation dynamique des réarrangements Vα-Jα du locus TRA/TRD chez la souris et chez l'homme.

Les réarrangements Vα-Jα constituent un mécanisme complexe car le locus TRA présente un nombre élevé de gènes, passe par plusieurs cycles de réarrangement, en absence d'exclusion allélique et contient le locus TRD [Krangel *et al* 2009]. Après la présentation des loci TRA/TRD humain et murin et des éléments Cis impliqués dans le contrôle des réarrangements Vα-Jα, les études dynamiques des réarrangements Vα-Jα chez la Souris et chez l'Homme seront abordés.

The Vα-Jα rearrangements constitute a complex mechanism, for the TRA locus presents a high number of genes, goes through multiple rearrangement rounds with no allelic exclusion, and encompasses the TRD locus [Krangel *et al* 2009]. After presenting the mice and humans TRA /TRD loci and the Cis elements involved in the control of Vα-Jα rearrangements, dynamical studies of Vα-Jα rearrangements in mice and humans will be addressed.

TRA/TRD locus in mice

In mice, the TRA locus, located on the chromosome 14 C1 at 19.7 cM spreads over 1650 kb. The locus of Balbc mice encompasses, from 5' to 3', 98 V genes, 60 J genes, and a single C gene. These genes are denoted respectively: TRAV, TRAJ and TRAC (Table 3). The mice TRD locus is embedded inside the TRA locus between TRAV and TRAJ genes. This mini locus presents, from 5' to 3', 5 TRDV, 2 TRDD, and 2 TRDJ genes, as well as a singe TRDC gene. Among TRAV genes, ten genes are also used for δ rearrangements and are called TRAV/DV. Another TRDV gene is situated in 3' of the TRDC gene, in inverted orientation. Hence, the locus contains a total of 16 TRDV (5+1 TRDV, 10 TRAV/DV). During evolution, the TRAV genes underwent duplications: small subsets of V genes were possibly duplicated first, followed by a more extended duplication, which gave both a Distal V-cluster and a Proximal V-Cluster (Figure 1) [Jouvin-Marche *et al*, 1989; Gahery-Segard *et al*, 1996; Glusman *et al*, 2001]. In mice, the majority of the TRAV genes are gathered in families: so, the 98 TRAV genes belong to 23

distinct families. The 60 TRAJ genes are spread over a small 64 kb region, thus the J region is zoomed in on Figure 1.

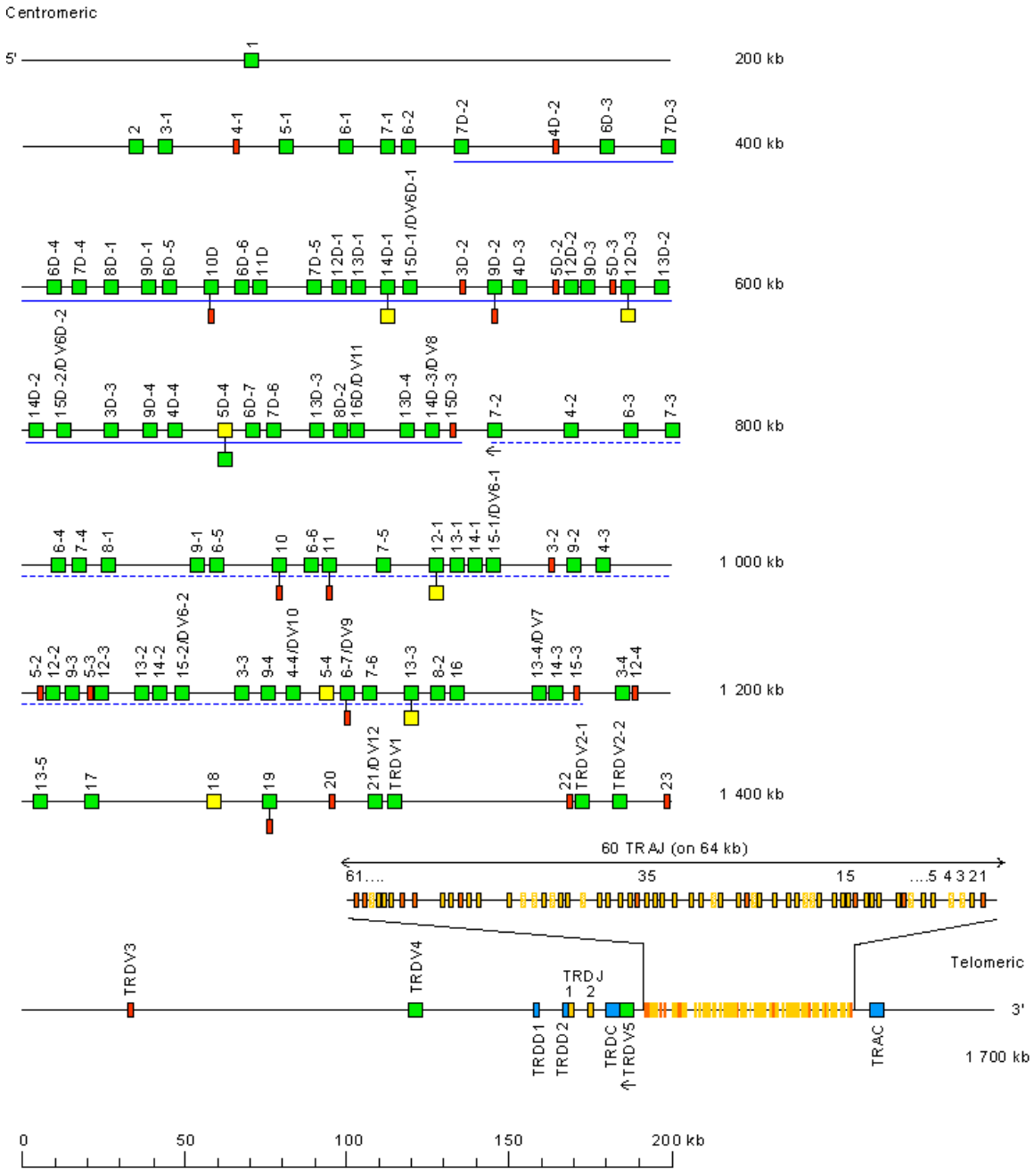


Figure 1. Mice TRA/TRD locus representation. From [Bosc *et al*, 2003] (representation available at www.imgt.org).

TRA/TRD locus in humans

The human TRA locus, located on chromosome 14 at 14q11.2, is spread over a 1000kb region. From 5' to 3', it is composed of 54 TRAV and 56 TRAJ genes, as well as one TRAC gene (Figure 2). The TRD locus is located inside the TRA locus, between the TRAV and TRAJ regions and presents a length of 60kb. This locus encompasses, from 5' to 3', 1 TRDV, 3 TRDD, 4 TRDJ genes and 1 TRDC gene. Five TRAV genes are also rearranged with TRDJ genes to generate the δ chain and are called TRAV/DV genes. One TRDV gene is situated among TRAV genes, at 360kb of the TRDC gene, another one is located in inverted orientation at the 3' side of the TRDC gene. Thereby, a total of 8 TRDV genes are present (1+1+1 TRDV, 5 TRAV/DV). In humans, the TRAJ region is composed of 61 genes and spread over 71kb.

The TRAJ regions are highly conserved between mouse and human, in terms of number of genes and length. More, the sequence of the genes and their functionality are relatively the same between the two species. Phylogenic analyses showed this TRAJ region homology is observed among mice, humans, and pigs as well, indicating a highly selective pressure for the conservation of the TRAJ genomic region throughout evolution [Uenishi *et al*, 2003]. The gene composition and length of the TRA/TRD loci in human and mouse are resumed on Table 1.

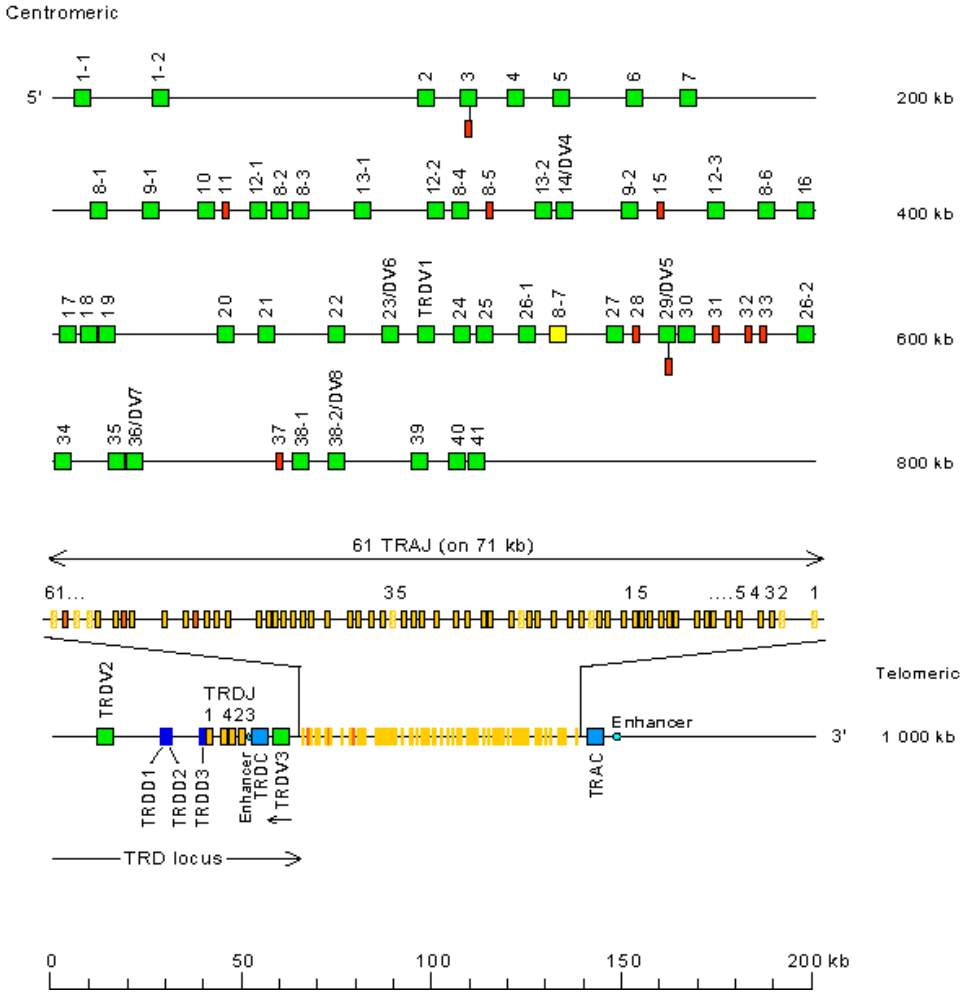


Figure 2. Human TRA/TRD locus representation. (representation available at www.imgt.org).

Locus	**TRA**		**TRD**	
Espèce	Humain	Souris	Humain	Souris
Taille du locus	1000Kb	1500Kb	60 kb	1500Kb
Nombre de gènes V	54	98	8 (3 TRDV + 5 TRAV/DV)	16 (6TRDV+10TRAV/DV)
Nombre de familles **V**	41	23	8	12
Nombre de gènes **D**	-	-	3	2
Nombre de gènes J	61	60	4	2
Nombre de gènes **C**	1	1	1	1

Table 1. Comparison of TRA/TRD locus gene number and length between mouse and humans.

TRA/TRD locus Cis regulating elements involved in the control of rearrangements

As described in the chapter one, the assembly of the V, (D), and J genes coding for the antigenic receptors is performed through a highly regulated recombination mechanism. In fact, V(D)J rearrangements occur at the different antigenic receptor loci exclusively in lymphocyte lineages and at precise stages of lymphocyte development [Yancopoulos and Alt, 1985 ; Cobb *et al*, 2006]. The specificity and temporal regulation of V(D)J rearrangements relies on the differential promoter activity of the RAG1/2 complex, on chromatin remodeling and on the nuclear relocation of the antigenic receptor loci. Each of the seven antigenic receptor loci includes Cis control elements, mostly enhancers and promoters, which target chromatin remodeling complexes and polymerase II binding, allowing the RAG complex to access its RSS substrates in a locus specific manner. The TRA/TRD Cis control elements involved in V(D)J rearrangement control are listed afterwards.

Enhancers δ and α

The compound TRA/TRD locus includes genes coding for the TRα and TRδ antigenic receptor chains. Dδ-Jδ and Vδ-DδJδ rearrangements occur during the DN-3 stage of T lymphocyte development whereas Vα-Jα rearrangements are first detected during the DN-4 stage in mice and mostly processed during the DP stage. This differential targeting of the V(D)J recombinase is directed by means of the δ and α Enhancers (Eα and Eδ). These enhancers promote chromatin remodeling by recruiting transcription factors and histone-modifying activities in a developmental stage-specific manner [McMurry and Krangel, 2000]. Enhancer functions also rely on their ability to achieve a subnuclear relocalisation of the associated antigenic receptor locus from nuclear periphery to transcription factories [Ragoczy *et al*, 2006]. Eα and Eδ are activated in a differential manner during T cell development; this switch, on and off, allows the distinctive rearrangement activities of the TRD and TRA locus during T cell development [Lauzurica and Krangel 1994; Hernandez-Munain *et al*, 1999].

Enhancer δ is located inside the TRD locus, between the TRDJ3 and TRDC genes in humans and between the TRDJ2 and TRDC genes in mice (Figure 3). The Enhancer δ favors transcriptional and recombination activities up to the TRDD1 gene, namely over a relatively

short distance of roughly 18Kb [Monroe *et al*, 1999]. The δ Enhancer position is crucial for its activity [Bassing *et al*, 2003].

Enhancer α is located at the 3' extremity of the TRAC gene, at the end of the TRA/TRD locus in mice and humans. This enhancer controls the TRAJ region accessibility and regulates the transcription rate of rearranged TRAV genes [Sleckman *et al*, 1998]. Eα activity spreads over a very extended distance, until 70kb. The functionality of Eα is ensured by a central 116pb region [Roberts *el al*, 1997]. Eα position inside the TRA locus is decisive for its activity [Bassing *et al*, 2003].

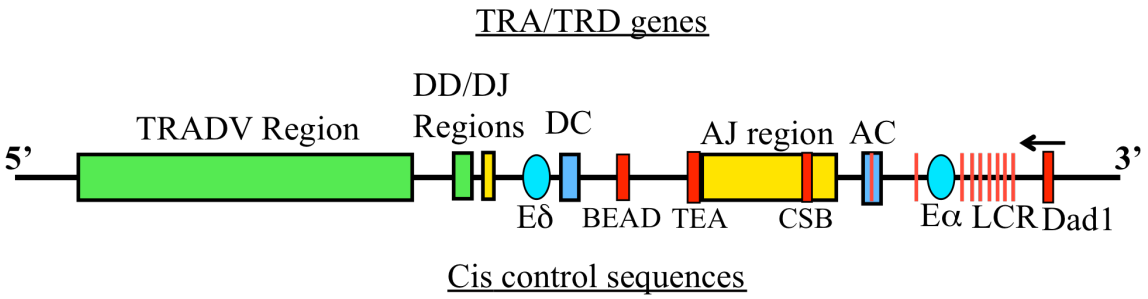


Figure 3. TRA/TRD locus organization and localization of the Cis regulatory sequences.

V genes promoters

Promoters correspond to regulatory sequences situated at the 5' end of genes, which allow transcription, the activation intensity being function of the distance between the promoter and the gene. In the antigenic receptor loci, each V segment embedded its own promoter; promoters are present as well in the 5' end of D and J genes. During DN stages of T cell development, no more than TRAV/DV and TRDV genes display a high transcriptional activity and are characterized by an accessible chromatin configuration [Hawwari and Krangel, 2005]. Each very promoter activity during the DN stages might define the V genes participating or not to the δ chains rearrangements. When the developing T cell enters the DP stage, numerous supplemental V genes are recently transcribed through their promoter activation by Eα. These newly accessible V genes will be available to participate to the α chain rearrangements [Krangel, 2009].

The TRA locus presents the particularity to undergo successive rearrangements: each Vα-Jα rearrangement event provoking a shortening of the locus length through the deletion of a T-cell Receptor Excision Circle (TREC). These successive contractions of the locus may facilitate or be facilitated by the V gene promoters / Eα interaction. This way, each shortening stage replaces V promoters, formerly located toward the 5' end of the locus, closer to the Eα,

allowing a progressive use of the V genes for rearrangements from the 3' to the 5' end of the TRAV region [Krangel, 2009].

J gene promoters: TEA and Jα49

The T-early α (TEA) and Jα49 sequences constitute two Jα promoters dependent on Eα activation. The TEA is located at the 5' end of the TRAJ region (Figure 1); it is conserved at 70% between human and mouse species. This Cis regulatory element controls germinal transcription and initiation of the rearrangements involving the 9 first J genes, located at the 5' end of the TRAJ region [Villey *et al*, 1996]. Accessibility of these first genes would be facilitated through germinal transcription: the histone acetylation and trimethylation (H3K4me3) coupled to the transcription would favor recombinase accessibility to J genes situated 12kb downstream of the TEA promoter [Abartegui and Krangel, 2007; Matthews *et al*, 2007]. On the contrary, the germinal transcription initiated from the TEA inhibits promoter activity of the J genes located from J47 to the 3' end of the TRAJ array through a presumed transcriptional interference process, which would inhibit chromatin accessibility [Abartegui and Krangel, 2006; Abartegui and Krangel, 2007].

In the absence of the TEA sequence (homozygote mutation TEA -/-), germinal transcripts initiated from the TRAJ 49 promoter are detected. The Jα49 promoter constitutes a second promoter Eα dependent, leading to a transcriptional activity on the TRAJ region [Hawwari *et al*, 2005].

Noticeably, a first Vα-Jα rearrangement inevitably deletes the TEA sequence and may equally take the Jα49 promoter away. Nevertheless, each rearrangement places the promoter of the rearranged V gene in the immediate 5' end proximity of the J genes that remains in the locus. The non-productive transcription initiated from this rearranged V promoter may be responsible of the chromatin remodeling implicated in the targeting of further rearrangements to the J genes situated immediately downstream the recent formed V-J association [Hawwari and Krangel, 2007]. This mechanism of successive contractions of the TRA locus places the new sites of transcription initiation continuously closer to the 3' end of the J array, allowing an efficient use the entire J region [Krangel, 2009].

Eventually, the TEA sequence is responsible of a synchronized usage of the J gene array between both the two alleles of each αβ T cell. As the TRA locus undergoes rearrangements with no genomic allelic exclusion, both the two TRA loci are rearranged in the majority of the cells; the two rearrangements involve J genes whose position in the germinal configuration is

close, mostly of less than 10 J genes. This interallelic correlation in the J usage is lost when the TEA is absent [Davodeau *et al.* 2001 ; Mauvieux *et al.* 2001]. No interallelic correlation is observed concerning the V usage.

BEAD-1: inhibits Eδ activity on the TRAJ region

The Eδ enhancer activates the TRDV genes situated at its 5' end, up to a 18kb distance. On its 3' end, Eδ activates the TRDV5 and TRDV3 genes respectively in mice and humans, located at a 10kb distance but does not activate TRAJ genes present only 4kb downstream. This specific activation is permitted by the Isolator Boundary Element Alpha Delta sequence (BEAD-1) found between TRDC gene and the TEA (Figure 1). BEAD-1 leads to a Cis inhibition of the Eδ activity [Zhong and Krangel 1997]. However, the deletion of BEAD-1 does not allow Eδ to induce Vδ-Jα rearrangements [Sleckman *et al*, 2001]. Thus, this restriction would also involve intrinsic TEA characteristics, which may prevent any TEA promoter activity during the DN stage of T cell development [Huang and Sleckman, 2007].

CSB of the TRAJ : spreads the Eα activity

The TRAJ region embeds a Conserved Sequence Block (CSB) of 125pb, located between the TRAJ4 and TRAJ3 genes. This sequence presents 95% of homology between mice and humans. The CSB binds nucleoprotein complexes that involve transcription factors whose expression is regulated throughout the T lymphopoiesis. A significant change in the CSB bound complex is observed during the DN and DP stage transition: the recognition of CSB by specific factors would be responsible of the Eα activation occurring during this stage transition. The CSB sequence would temporarily regulate the chromatin structure and would relay Eα activity to make its activity efficient over the whole 70kb TRAJ region [Chen and Kuo 2001].

Locus Control Region

In a general point of view, the developmental and cell lineage-specific regulation of gene expression relies on various Cis acting regulatory elements (enhancers, promoters, silencers), which include Locus Control Regions (LCR). A Locus Control Region corresponds to a set of regulatory sequences defined by their ability to enhance the expression of the linked genes up to the physiological expression levels.

LCR was first identified in the human β-globin locus: it was determined to be an essential element for the normal regulation of beta-globin gene expression [Grosveld *et al*, 1987; Gerstein

et al, 2007]. A LCR contains a set of regulatory elements, which once integrated in a heterologue transgene, conserves the regulatory skills of the region from which it originates [Greaves, 1989]. These regulatory features include the tissue and developmental stage specificities as well as the linked gene expression level. The capacity of a LCR to regulate a gene transcription level indicates its ability to protect a transgene from the action of other regulatory elements (silencer, enhancers). Consequently, LCR is able to isolate a gene from its position effects: position effects occuring when Cis regulatory elements alter the pattern or level of gene expression expected from a particular transgene [Palmiter *et al*, 1986].

The TRA/TRD locus contains a LCR located at its 3' extremity. The LCR spreads over 5kb, from the TRAC gene second exon to the Dad1 gene, a ubiquitously expressed anti-apoptotic gene whose deletion is lethal (Figure 1) [Nakashima *et al*, 1993]. The very close linkage among TRA, TRD, and Dad1 is evolutionarily conserved [Wang *et al*, 1997]. Evidently, the achievement of proper and individual regulations for the TRA, TRD, and Dad1 genes, located within the same locus constitutes a complex assignment [Krangel *et al*, 2004]. TRA or TRG genes are somatically rearranged during T cell development and afterwards only expressed in αβ and δγ T cells respectively whereas the Dad1 gene is ubiquitously expressed [Hong *et al*, 1997].

The TRA/TRD LCR presents DNase-I hypersensitive sites (HS), from HS1, HS1', HS2 through HS6. The different hypersensitive sites do not support the same regulatory functions: HS1 sustains a transcriptional enhancer activity [Winoto *et al*, 1989]; HS1', located downstream Eα, provides the tissue specificity [Ortiz *et al*, 1999]; HS1', HS2 through HS6 are critical for the ability of the LCR to protect transgenes from position effects [Diaz *et al*, 1994; Ortiz *et al*, 1999], and finally HS6 would be involved in the commutation between Dad1 transcriptional activity and the lineage specific activation of TRAD rearrangements [Ortiz *et al*, 2001]. Independently, these TRA/TRD LCR sub-elements may play a role in managing the separate regulatory programs of the TCR and Dad1 genes, thus preventing their inappropriate interaction during T cell development. In this way, some of the LCR sub-elements might act as insulator/boundary-like elements. Finally, concerning the TRA/TRD locus rearrangements, the LCR affects long-range chromatin structure and epigenetic modifications [Ortiz *et al*, 1999; Santoso *et al*, 2000].

Dynamical Modeling of TRA/TRD Vα-Jα Rearrangements in Mice

The TR repertoire constitutes a key element of the specific immune system. Consequently, the knowledge about the dynamical building of the TCRαβ repertoire and the evaluation of its diversity present major medical interests. Formerly, the TCRαβ repertoire evaluation was based on both the TCRβ chain analysis through a cellular approach (flux cytometry) and the CDR3 analysis using molecular methods. The study of the TCRα chains represents a more difficult task because of the complexity of both the TRA locus structure and the Vα-Jα successive rearrangement mechanism. However, investigations on TCRβ chain only provide a partial sight of the TCRαβ repertoire, and more, the TCRα chain may play a predominant role in the recognition of the CMH/peptide complex by the TCR. The biomodeling study of the Vα-Jα successive rearrangement process in mice, based on extensive experimental quantifications, offered a clear understanding of the mouse α repertoire dynamical building and proposed a simulated repertoire showing the frequencies of the entire V-J associations.

Experimental Background

The dynamic modeling addressed the gene use during successive Vα-Jα rearrangements in mice, and was based on experimental quantifications of particular Vα-Jα associations. The article 1, fully presented in the annex 1 of the present chapter, addressed the determination of Vα-Jα association status:

Quantitative and qualitative changes in V-J alpha rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor alpha chain repertoire. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche PN, Jouvin-Marche E. J Exp Med. 2002 Nov 4;196(9):1163-73.

In this study, the use of multiplex genomic PCR assays and real time PCR analyses gave a reliable drawing of the TRα thymic combinatorial repertoire shape in mice. Results determined that the relative frequencies of the particular Vα-Jα associations studied were dependent on the position of the Vα and Jα genes in the locus germline configuration. These analyses confirmed previous study results, which showed the Jα genes were used in a progressive manner through the rearrangement process, from the TRAJ region 5' end toward the 3' end [Thompson *et al*, 1990; Rytkonen *et al*, 1996]. Concerning the TRAV region use, little knowledge was available: a previous study determined the occurrence of differential transcription activities among the

ADV2 family genes during ontogeny, but the experiments were performed at the transcript level. In order to get away from transcriptional biases, the study reported in the article 1 was performed directly at the genomic level. Eventually, this work established the TRA gene recombination mechanism would follow a progressive use of the V and J regions, each given V studied associating with not all but a given subset of subsequent J genes, these preferential associations restricting the potential TRα combinatorial repertoire.

Gene data easier access for rearrangement studies: the IMGT/TCRGeneInfo tool

The interactive online information system IMGT/TCRGeneInfo was developed in order to facilitate the access to the TRA/TRD and TRB loci gene data in human and mouse species for rearrangement-oriented tasks. Indeed, all the information of interest concerning rearrangements is gathered: after the choice of a couple of genes of interest, the respective RSS sequences, positions of the genes in the germline configuration, and sequences of the genes appeared on a single page along with a simplified locus representation (Figure 4). A comprehensive description of the IMGT/TCRGeneInfo system is available in the article 2 (annex 2 of the present chapter):
IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Thierry-Pascal Baum, Nicolas Pasqual, Florence Thuderoz, Vivien Hierle, Denys Chaume, Marie-Paule Lefranc, Evelyne Jouvin-Marche, Patrice-NoEl Marche and Jacques Demongeot. Nucleic Acids Research, 2004, Vol. 32:51-54

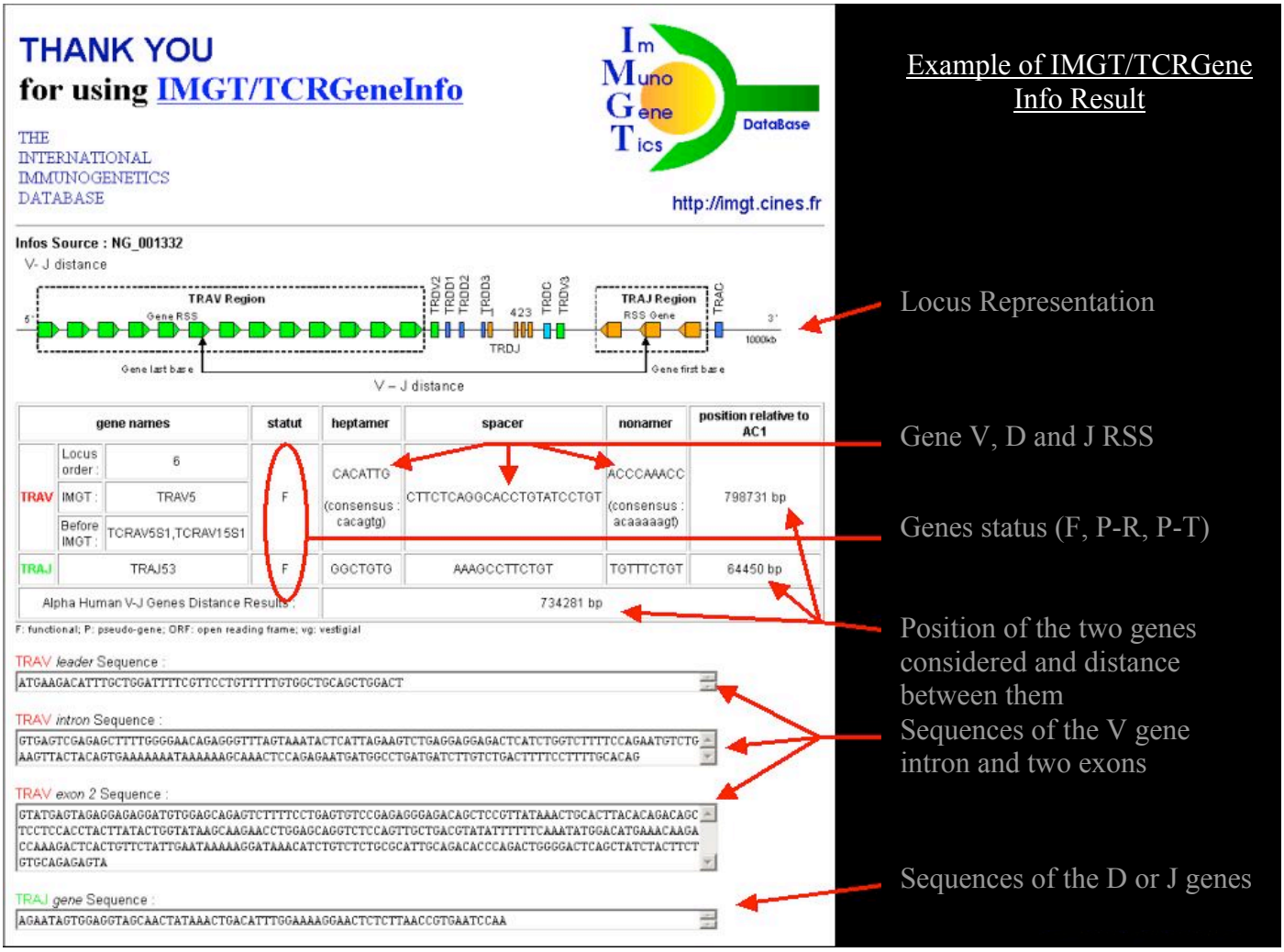


Figure 4. Screenshot of a GeneInfo result page. From top to bottom are displayed: a locus representation showing the rearrangement studied; a table gathering the nomenclature of the selectioned genes; the RSS sequences along with the consensus sequence from [Glusman *et al*, 2001]. The gene position in the locus is calculated from the first base of the TRAC first exon. Each sequence is available for the users (copy/paste). This information is accessible for each of the 4 TR chains from the IMGT database (www.imgt.org).

Previous models concerning gene segment use in rearrangements

Biomodeling studies addressing the gene usage in V(D)J rearrangements were performed for the IgK locus in B cells. This locus contains only 4 functional Jk genes segments used with a slight bias. From these studies, a quasi-sequential use of these genes was found to fit the best the available experimental Jk data. The models developed either assigned probabilities to each of the four genes [Mehr *et al*, 1999], or varied the ratio of the likelihood ratio of choosing each gene segment and the one immediately upstream [Louzoun *et al*, 2002].

Regarding the gene use in Vα-Jα rearrangements, Warmflash and Dinner published a first model of TRAD locus recombinations [Warmflash *et al*, 2006], using thymic data from our article 1 [Pasqual *et al*, 2002]. In this model, two accessibility windows move over the V and J regions, allowing rearrangements. The windows sizes were theoretically estimated from the average separation of the gene segments on the different TRA alleles. This model was validated

using the thymic relative quantifications of a set of 8 J genes by two V genes (Vα6 and Vα19). This first model was not able to account for the progressive transformation of the frequency curves of the J region use by the V genes, from the proximal V to the distal ones. In particular, the occurrence of a bimodal distribution of these frequencies for the intermediate V genes did not emerge from the simulations.

A brownian Rachet Model

PCR quantifications of particular V-J rearrangements from Balb-c mouse thymic extracts determined the J region use by V genes described by Poisson-like as well as Gaussian-like or Gaussian mixture distribution curves (Article 1). In order to state if a simple probabilistic model, related to a pure random mechanism, would fit these curves, a model based on the use of the Brownian Motor Theory was firstly developed. In Fact, back in time, the DNA loops were successfully modelized with the use of the Brownian Motor Theory, also called Worm-like Chain Theory [Merlitz *et al*, 1998]. According to this theory, the DNA chain is described as rigid units, interacting by harmonic bending, twisting, and stretching potentials. This DNA mathematical representation gave valuable explanations of the structure and dynamics of linear and superhelical DNAs, the Monte Carlo procedure being used to calculate thermodynamic equilibrium of the structural ensembles [Rybenkov *et al*, 1997]. In a Brownian ratchet model accounting for the loop formation during V-J rearrangements, the double stranded string of DNA would be considered to fold and form a loop with two strands crossing each other, the V and J regions crossing each other at a point called *P*. After elimination of the intermediary Delta region (first maturation), one part of the string would correspond to the V locus and the other to the J one (Figure 2), the V and J regions crossing each other at the called the *P* point. The V and J loci would move under random constraints like thermal agitation, sliding through the *pivot P*, and eventually, each given rearrangement would be made of the very V and J genes located at the *P* point. The constraints on the performed sliding being described “click by click”, according to the Brownian Motor Theory, a click corresponding to a unit of displacement along one of the V or J regions. Each displacement unit is supposed to be a V or a J gene. Thus, the Pivot point can move along the V locus from the most proximal V genes (*Vprox*) to the most distal ones (*Vdist*) with 104 possible consecutive clicks corresponding to the 104 V genes on the V region in Balb-c mice; the same *P* displacements being achievable over the 49 genes of the J region. The probabilities of displacement of the *P* point towards the distal or proximal V genes are respectively called *pVdist* and *pVprox*. *Vdist* denotes the distance between *Vdist* and *P*, and

Vprox the distance between *Vprox* and *P* (idem for the J region). Eventually, V denotes the position of *P* along the V region string, and J the position of *P* along the J region string. Biological results, demonstrated that proximal V and J genes are used preferentially compared to distal V and J ones. Thus, it can be assumed that the displacement of the pivot *P* is proportional to *Vdist* and *Jdist* distances, and inversely proportional to *Vprox* and *Jprox* ones. Considering the displacement of *P* as random, and denoting α the proportionality factor ensuring that the sum of the considered probabilities is normalized to 1, the following equations come:

pVprox is proportional to *Jdist* and *Vdist*:

$$\boldsymbol{pVprox} = \alpha * Jdist * Vdist$$

pVdist is proportional to *Jprox* and *Vprox*:

$$\boldsymbol{pVdist} = \alpha * Jprox * Vprox$$

$$\boldsymbol{pVprox + pVdist} = 1 = \alpha * [(Jdist*Vdist) + (Jprox*Vprox)]$$

From these relations the following equations can be inferred:

$$\boldsymbol{pVprox(k)}=(Vdist(k-1)*Jdist(k-1))/((Jprox(k-1)*Vprox(k-1)) + (Vdist(k-1)*Jdist(k-1)))$$

$$\boldsymbol{pVdist(k)}=(Vprox(k-1)*Jprox(k-1)) / ((Jprox(k-1)*Vprox(k-1)) + (Vdist(k-1)*Jdist(k-1)))$$

$$pJprox(k) = pVprox(k) \text{ and } pJdist(k) = pVdist(k)$$

From these equations, is applied the following algorithm for the displacement of *P*:

1) for the initial conditions, are randomly chosen:

- V: initial position of *P* on the V strand
- J: initial position of *P* on the J strand

2) for each Brownian click, 2 random choices are made by using 2 random variables *T1* and *T2*, supposed to be uniform on [0,1]:

- *T1* for *pVprox*:
  - If $T1 \leq pVprox$: *P* is moved by 1 click towards *Vprox*
  - If $T1 > pVprox$: *P* is moved by 1 click towards *Vdist*
- *T2* for *pJprox*:
  - If $T2 \leq pJprox$: *P* is moved by 1 click towards *Jprox*
  - If $T2 > pJprox$: *P* is moved by 1 click towards *Jdist*

T1 and *T2* choices are repeated 3, 10 and 50 times corresponding to 3, 10 and 50 Brownian clicks. The position of *P* after the clicks series gives the final rearrangement position. A set of 10^6 TRA loci performing rearrangements is simulated, and the J region uses by V1 and V1 in the simulated population are plotted on the graphs presented on the Figure 3.

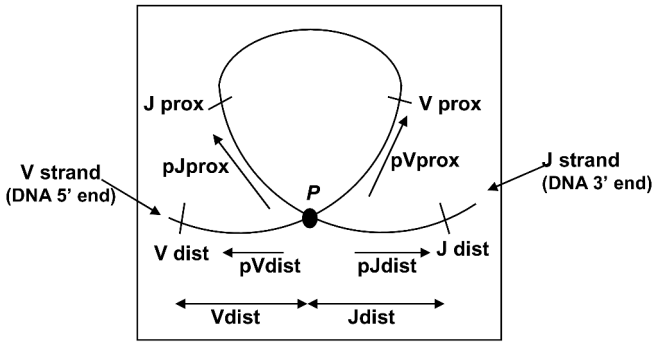


Figure 2. V and J loci crossing each other at the *P* point in order to realize rearrangements after the deletion of the TRD locus.

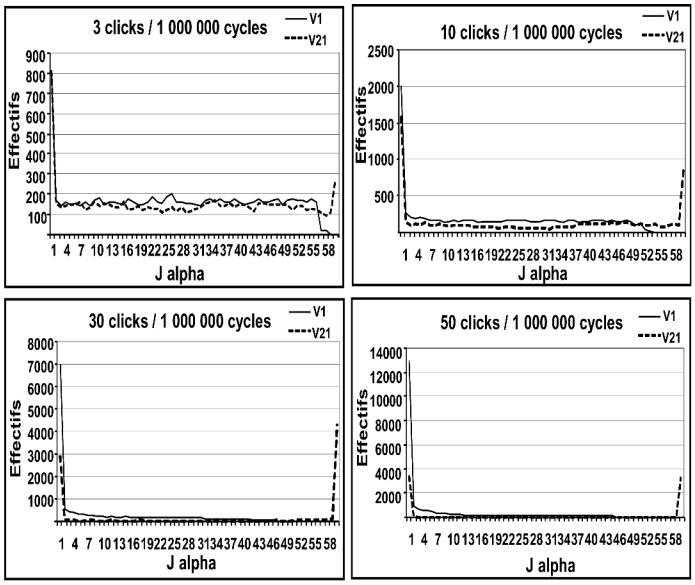


Figure 3. Simulation rearrangement frequencies obtained for an increasing number of clicks (3 to 50) in the Brownian ratchet model.

From the curves of Figure 3, obtained by performing successively 3, 10, 30 and 50 clicks, it can be deduced that over 30 clicks, there is no more evolution, the rearrangement being fixed. The numerical curve slopes resulting from simulations are somehow random and do not present the typical regularity of the experimental curve slopes (Article 1). Thus, the Brownian ratchet model is unable to explain experimental curves, which allows making the statement than pure random phenomenon is definitely not able to originate the experimental distributions observed.

Successive windowing model

The stochastic successive model we developed for the Vα-Jα rearrangements of the TRA/TRD locus in mice (Balbc) brought new insights on the mechanism dynamical knowledge for it was based on experimentally estimated parameters. Indeed, it included the exact physical

position of the Vα and Jα genes in the TRA/TRD locus and two opening speed intervals for the V and J regions determined from ontogeny experimental results. The overall delay for the whole rearrangement process was determined from experiments as well. The model-simulated results were validated with thymic data and afterwards with peripheral data, constituted by the Jα region use by the 5 TRAV21 family members. This study is comprehensively presented in the article 3 (annex 3):

Numerical Modelling Of The V-J Combinations Of The T Cell Receptor TRA/TRD Locus. Thuderoz F, Simonet MA, Hansen O, Pasqual N, Dariz A, Baum TP, Hierle V, Demongeot J, Marche PN, Jouvin-Marche E. PLoS Computational Biology 2010 Vol 6. e1000682. 1-12.

Dynamical Modeling of TRA/TRD Vα-Jα Rearrangements in Human

In the transition from the mouse model to human, an extensive experimental study was performed at the thymic level, using real time PCRs. In spite of the human TRA gene polymorphism, the quantifications from thymic genomic DNA displayed a good inter-individual consistence, offering a first experimental wide-ranging sampling of the human TRα repertoire in the thymus. In this study, the successive windowing model was adapted to the human locus structure particularities and the simulation results showed a good coherence with both our experimental V-J quantifications and with the interallelic J usage from [Davodeau *et al*, 2001]. The human study is presented in the article 4 (annex 4):

Numerical Model for the Vα-Jα Gene Use in Human TRA/TRD locus: Recombination Dynamical Rules. Thuderoz F, Hansen O, Simonet MA, Daris A, Borel E, Marche PN , Jouvin-Marche E, Demongeot J. (submitted).

Perspectives concerning the Vα-Jα Rearrangement biomodeling

Toward a characterization of the immune response: identification of bias in the combinatorial TR α repertoire

The knowledge about the human thymic repertoire shape constitutes by itself a key issue in the immune system development and a crucial requirement in therapeutic interventions aiming to reconstitute or to control immune responses. The next step in this combinatorial repertoire study will be to come back to the mouse model in order to characterize the immune response based on the established knowledge about the primary repertoire. This research would require timed quantifications of numerous V-J associations after a chosen antigen contact. Large statistical analyses would be crucial in this perspective study in order to define the V-J association frequency values being inside normal inter-individual variations or constituting a component of the immune response.

Links toward a model based on locus distinct topologies

Recently, new advances were achieved concerning the IgH locus topology variations during the development of B lineage cells. *In vitro* experiments showed the IgH locus contractions bring V_H genes into a relative close physical proximity to the D_H-J_H segments, previously to VDJ gene rearrangements. These variations would lead to the formation of

configurations in clusters of loops, proposed to allow long-range genomic interactions to occur with relative high frequency [Jhunjhunwala *et al*, 2009]. Concerning the TRA locus, the recombination centers initiated over the J region appeared to recruit the second RSS to form the synaptic complex [Ji *et al*, 2010]. In addition, the opening speeds of the V and J regions determined experimentally through the first days of detection of the rearranged genes were satisfactorily incorporated in a stochastic successive windowing model to promote dynamical explanation of the Vα-Jα rearrangement process in mice, and afterwards in humans (Articles 3 and 4). Therefore, the progression of the accessibility windows over the V and J regions would be to consider in terms of loop formation possibilities, the window progressions standing for the control exerted on the structure of loops of a given size and a given positioning variability throughout the successive rearrangement process. Experiments determining the TRA topology evolution throughout rearrangements would be interestingly compared with the window accessibility progression interval speeds; however, at the present time, the opening speeds determined from biological data correspond to the most available valuable and comprehensive mechanistic explanation of observed V-J association frequencies.

References of the Chapter III

Abarrategui I, Krangel MS. Noncoding transcription controls downstream promoters to regulate T-cell receptor alpha recombination. EMBO J. 2007. 26:4380-90.

Abarrategui I, Krangel MS. Regulation of T cell receptor-alpha gene recombination by transcription. Nat Immunol. 2006. 7:1109-15.

Bassing CH, Tillman RE, et al. T cell receptor (TCR) alpha/delta locus enhancer identity and position are critical for the assembly of TCR delta and alpha variable region genes. Proc Natl Acad Sci U S A. 2003. 100: 2598-603.

Bosc N, Lefranc MP. The mouse (Mus musculus) T cell receptor alpha (TRA) and delta (TRD) variable genes. Dev. Comp. Immunol. 2003. 27:465-497.

Chen ML, Kuo CL. A conserved sequence block in the murine and human T cell receptor Jalpha loci interacts with developmentally regulated nucleoprotein complexes in vitro and associates with GATA-3 and octamer-binding factors in vivo. Eur J Immunol. 2001. 31: 1696-705.

Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM. Accessibility control of V(D)J recombination. Adv Immunol. 2006. 91:45-109.

Davodeau F, Difilippantonio M, et al. The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. Embo J. 2001. 20: 4717-29.

Diaz P, Cado D, Winoto A. A locus control region in the T cell receptor / locus. Immunity. 1994.1:207-217.

Gahery-Segard, H., E. Jouvin-Marche, et al. Germline genomic structure of the B10.A mouse Tcra-V2 gene subfamily. Immunogenetics. 1996. 44: 298-305.

Gerstein MB, Bruce C, Rozowsky JS, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007. 17:669–81.

Glusman, G., L. Rowen, et al. Comparative genomics of the human and mouse T cell receptor loci. Immunity. 2001. 15: 337-49.

Greaves DR, Antoniou M, van Assendelft GB, Collis P, Dillon N, Hanscombe O, Hurst J, Lindenbaum M, Talbot D, Grosveld F. The beta-globin dominant control region. Prog Clin Biol Res. 1989. 316A:37-46.

Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell. 1987;51:975-985

Hawwari A, Krangel MS. Role for rearranged variable gene segments in directing secondary T cell receptor alpha recombination. Proc Natl Acad Sci U S A. 2007. 104:903-7.

Hawwari A, Bock C, Krangel MS. Regulation of T cell receptor alpha gene assembly by a complex hierarchy of germline Jalpha promoters. Nat Immunol. 2005. 6:481-9.

Hawwari A, Krangel MS. Regulation of TCR delta and alpha repertoires by local and long-distance control of variable gene segment chromatin structure. J Exp Med. 2005. 202:467-72.

Hernandez-Munain C, Sleckman BP, et al. A developmental switch from TCR delta enhancer to TCR alpha enhancer function during thymocyte maturation. Immunity. 1999. 10: 723-33.

Hong NA, Cado D, Mitchell J, Ortiz BD, Hsieh SN, Winoto A. A targeted mutation at the T-cell receptor / locus impairs T-cell development and reveals the presence of the nearby antiapoptosis gene Dad1. Mol. Cell. Biol. 1997. 17:2151-2157.

Huang CY, Sleckman BP. Developmental stage-specific regulation of TCR-alpha-chain gene assembly by intrinsic features of the TEA promoter. J Immunol. 2007 179:449-54.

Jhunjhunwala S, van Zelm MC, Peak MM, Murre C. Chromatin architecture and the generation of antigen receptor diversity. Cell. 2009.138:435-48.

Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell. 2010.141:419-31.

Jouvin-Marche, E., M. G. Morgado, et al. Complexity, polymorphism, and recombination of mouse T-cell receptor alpha gene families. Immunogenetics. 1989. 30: 99-104.

Krangel MS. Mechanics of T cell receptor gene rearrangement. Curr Opin Immunol. 2009 21:133-9.

Krangel MS, Carabana J, Abbarategui I, Schlimgen R, Hawwari A. Enforcing order within a complex locus: current perspectives on the control of V(D)J recombination at the murine T-cell receptor / locus. Immunol. Rev. 2004. 200:224-232.

Lauzurica P, Krangel MS. Temporal and lineage-specific control of T cell receptor alpha/delta gene rearrangement by T cell receptor alpha and delta enhancers. J Exp Med. 1994. 179:1913-21.

Louzoun, Y., T. Friedman, E. L. Prak, S. Litwin, and M. Weigert. 2002. Analysis of B cell receptor production and rearrangement part I. Light chain rearrangement. Semin. Immunol. 14: 169–190.

Matthews AG, Kuo AJ, Ramón-Maiques S, Han S, Champagne KS, Ivanov D, Gallardo M, Carney D, Cheung P, Ciccone DN, Walter KL, Utz PJ, Shi Y, Kutateladze TG, Yang W, Gozani O, Oettinger MA. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature. 2007. 450:1106-10.

Mauvieux L, Villey I, et al. T early alpha (TEA) regulates initial TCRVAJA rearrangements and leads to TCRJA coincidence. Eur J Immunol. 2001. 31: 2080-6.

McMurry MT, Krangel MS. A role for histone acetylation in the developmental regulation of VDJ recombination. Science. 2000. 287: 495-8.

Mehr, R., M. Shannon, and S. Litwin. 1999. Models for antigen receptor gene rearrangement. I. Biased receptor editing in B cells: implications for allelic exclusion. J. Immunol. 163: 1793–1798.

Merlitz H, Rippe K, Klenin KV, Langowski J. Looping dynamics of linear DNA molecules and the effect of DNA curvature: a study by Brownian dynamics simulation. Biophys J. 1998. 74:773-9.

Monroe RJ, Sleckman BP, et al. Developmental regulation of TCR delta locus accessibility and expression by the TCR delta enhancer. Immunity. 1999. 10: 503-13.

Nakashima, T., T. Sekiguchi, A. Kuraoka, K. Fukushima, Y. Shibata, S. Komiyama, T. Nishimoto. Molecular cloning of a human cDNA encoding a novel protein, DAD1, whose defect causes apoptotic cell death in hamster BHK21 cells. Mol. Cell. Biol. 1993. 13: 6367-6374.

Ortiz BD, Harrow F, et al. Function and factor interactions of a locus control region element in the mouse T cell receptor-alpha/Dad1 gene locus. J Immunol. 2001. 167:3836-45.

Ortiz BD, Cado D, Winoto A. A new element within the T-cell receptor locus required for tissue-specific locus control region activity. Mol. Cell. Biol. 1999. 19:1901-1909.

Palmiter, R. D., R. L. Brinster. Germ-line transformation of mice. Annu. Rev. Genet. 1986. 20: 465-499.

Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche PN, Jouvin-Marche E. Quantitative and Qualitative Changes in ADV-AJ Rearrangements During Mouse Thymocytes Differentiation : Implication For a Limited TCR ALPHA Chain Repertoire. J. Exper. Medicine. 2002. 196:1163-1174.

Ragoczy T, Bender MA, Telling A, Byron R, Groudine M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. Genes Dev. 2006. 20:1447-57.

Roberts JL, Lauzurica P, Krangel MS. Developmental regulation of VDJ recombination by the core fragment of the T cell receptor alpha enhancer. J Exp Med. 1997. 185:131-40.

Rybenkov VV, Vologodskii AV, Cozzarelli NR. The effect of ionic conditions on DNA helical repeat, effective diameter and free energy of supercoiling. Nucleic Acids Res. 1997. 25:1412–1418.

Rytkonen MA, Hurwitz JL, et al. Restricted onset of T cell receptor alpha gene rearrangement in fetal and neonatal thymocytes. Eur J Immunol. 1996. 26:1892-6.

Santoso B, Ortiz BD, Winoto A. Control of organ-specific demethylation by an element of the T-cell receptor- locus control region. J. Biol. Chem. 2000. 275:1952-1958.
Thompson SD, Pelkonen J, et al. First T cell receptor alpha gene rearrangements during T cell ontogeny skew to the 5' region of the J alpha locus. J Immunol. 1990. 145: 2347-52.

Thompson SD, Pelkonen J, et al. Nonrandom rearrangement of T cell receptor J alpha genes in bone marrow T cell differentiation cultures. J Immunol. 1990. 144:2829-34.

Thuderoz F, Simonet MA, Hansen O, Pasqual N, Dariz A, Baum TP, Hierle V, Demongeot J, Marche PN, Jouvin-Marche E. Numerical Modelling Of The V-J Combinations Of The T Cell Receptor TRA/TRD Locus. PLoS Computational Biology. 2010. e1000682:1-12.

Villey I, Caillol D, et al. Defect in rearrangement of the most 5' TCR-J alpha following targeted deletion of T early alpha (TEA): implications for TCR alpha locus accessibility. Immunity. 1996. 5: 331-42.

Sleckman BP, Carabana J, et al. Assessing a role for enhancer-blocking activity in gene regulation within the murine T-cell receptor alpha/delta locus. Immunology. 2001. 104: 11-8.

Sleckman BP, Bassing CH, Bardon CG, Okada A, Khor B, Bories JC, Monroe R, Alt FW. Accessibility control of variable region gene assembly during T-cell development. Immunol Rev. 1998.165:121-30.

Uenishi, H., H. Hiraiwa, et al. Genomic structure around joining segments and constant regions of swine T-cell receptor alpha/delta (TRA/TRD) locus. Immunology. 2003. 109: 515-26.

Wang K, Gan L, Kuo CL, Hood L. A highly conserved apoptotic suppressor gene is located near the chicken T-cell receptor chain constant region. Immunogenetics. 1997. 46:376-382.

Warmflash A, Dinner AR. A model for TCR gene segment use. J Immunol. 2006. 177:3857-3864.

Winoto A, Baltimore D. A novel, inducible and T cell-specific enhancer located at the 3' end of the T cell receptor locus. EMBO J. 1989. 8:729-733.

Yancopoulos GD, Alt FW. Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell. 1985. 40: 271-81.

Zhong XP, Krangel MS. An enhancer-blocking element between alpha and delta gene segments within the human T cell receptor alpha/delta locus. Proc Natl Acad Sci USA. 1997. 94: 5219-24.

Annex 1

Quantitative and qualitative changes in V-J alpha rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor alpha chain repertoire. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche PN, Jouvin-Marche E. J Exp Med. 2002 Nov 4;196(9):1163-73.

Quantitative and Qualitative Changes in V-J α Rearrangements During Mouse Thymocytes Differentiation: Implication For a Limited T Cell Receptor α Chain Repertoire

Nicolas Pasqual,[1] Maighréad Gallagher,[1] Catherine Aude-Garcia,[1] Mélanie Loiodice,[1] Florence Thuderoz,[2] Jacques Demongeot,[2] Rod Ceredig,[1] Patrice Noël Marche,[1] and Evelyne Jouvin-Marche[1]

[1]*Laboratoire d'Immunochimie, Commissariat à l'Energie Atomique, Institut National de la Santé et de la Recherche Médicale, Unité 548, Université Joseph Fourier, 38054 Grenoble Cedex 9, France*

[2]*Technique pour l'Imagerie, la Modélisation et la Cognition, Université J Fourier Grenoble, Faculté de Médecine, 38700 la Tronche, France*

Abstract

Knowledge of the complete nucleotide sequence of the mouse TCRAD locus allows an accurate determination V-J rearrangement status. Using multiplex genomic PCR assays and real time PCR analysis, we report a comprehensive and systematic analysis of the V-J recombination of TCR α chain in normal mouse thymocytes during development. These respective qualitative and quantitative approaches give rise to four major points describing the control of gene rearrangements. (a) The V-J recombination pattern is not random during ontogeny and generates a limited TCR α repertoire; (b) V-J rearrangement control is intrinsic to the thymus; (c) each V gene rearranges to a set of contiguous J segments with a gaussian-like frequency; (d) there are more rearrangements involving V genes at the 3′ side than 5′ end of V region. Taken together, this reflects a preferential association of V and J gene segments according to their respective positions in the locus, indicating that accessibility of both V and J regions is coordinately regulated, but in different ways. These results provide a new insight into TCR α repertoire size and suggest a scenario for V usage during differentiation.

Key words: TCR-diversity • TCRAD locus • rearrangement • development • quantitative PCR

Introduction

Mature T lymphocytes express a clonotypic TCR on their surface. The TCR is activated through recognition of an antigenic peptide presented by molecules of the major histocompatibility complex. The TCR is composed of αβ or γδ heterodimers, in which each chain consists of a variable and constant region, in association with the CD3 complex composed of ε, γ, δ, and ζ chains (1). During T cell development, TCR α, β, γ, and δ chains are assembled following the rearrangement of independent gene segments contained on the corresponding TCRAD (coding TCRα and TCRδ), TCRB, and TCRG loci. This rearrangement process uses an enzymatic complex (the V(D)J recombinase) that selectively targets recombination signal sequences (RSS)* flanking the coding sequences of dispersed V, D (only for TCRB and TCRD), and J gene segments (2, 3). In addition to the combinatorial process, TCR repertoire diversity is enhanced by various mechanisms which include imprecise joining at V(D)J junctions, addition or removal of junctional nucleotides, and pairing constraints of different α and β (or γ and δ) chain molecules (4, 5). V(D)J recombination is a highly regulated process in terms of both cell lineage and stage of T cell development (6). TCRB rearrangement is initiated at the $CD4^{-}CD8^{-}$ double-negative stage (7). The expression of a productively rearranged TCRβ protein associated with the invariant pTα chain and

M. Gallagher's present address is Lymphocyte Activation Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 61 Lincoln's Inn Fields, London WC2A 3PX, UK.

Address correspondence to Dr. Evelyne Jouvin-Marche, Laboratoire d'Immunochimie CEA-G/DRDC/ICH, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France. Phone: 33-4-3878-5770; Fax: 33-4-3878-9803; E-mail: immuno@dsvsud.cea.fr

**Abbreviations used in this paper:* DP, double positive; FTOC, fetal thymic organ culture; IMGT, ImMunoGeneTics database; RSS, recombination signal sequence.

 J. Exp. Med. © The Rockefeller University Press • 0022-1007/2002/11/1163/11 $5.00
Volume 196, Number 9, November 4, 2002 1163–1173
http://www.jem.org/cgi/doi/10.1084/jem.20021074

the CD3 complex leads to progression to the $CD4^+CD8^+$ double-positive (DP) stage and the initiation of rearrangement at the TCRAD locus (for a review, see reference 8).

Recently, our laboratory has established a detailed map of the BALB/c mouse TCRAD locus (9). This map tallies closely with a map of the TCRAD locus derived from the 129/SvJ mouse strain which has recently been constructed (NCBI accession nos.: AE008683-AE008686). Based on restriction length polymorphism it has been found that BALB/c and 129/SvJ mice strains share the same TCRA haplotype (10). Sequence alignments of the BALB/c and 129/SvJ TCRAD loci reveal that the level of sequence polymorphism is very low and that the distribution of V segments is identical (unpublished data). For instance, the majority of sequences from numerous TCRA haplotypes have been found to be closely related with few nucleotide substitutions (11, 12). Sequencing analysis of alleles of V genes belonging to the same V family reveals that the polymorphisms between TCRA haplotypes arose by duplication or deletion events which result in a gain or loss of V segments. Thus, as for human immunoglobulin VH gene polymorphism, allelic differences are likely to have little influence on the shape of the repertoire (13).

BALB/c and 129/SvJ TCRAD loci are 1,200 kb in length and contain 104 V and 60 J segments. The D, J, and C genes encoding δ chains are found between the V and J α segments. The V elements can be organized into 23 α and 6 delta families in which different members of the same family are at least 75% identical at the nucleotide level. Members of the same family are interspersed along the locus. Interestingly, single-membered families (e.g. V12, V19, V6) tend to localize at the extremities of the V region. With the exception of one member of the V1 family, the orientation of all segments is consistent with V-J joining occurring by a deletional mechanism. 18 V segments have been described as pseudogenes due to either the absence of an open reading frame (ORF) or because they encode a polypeptide defective on structural grounds (i.e., unable to fold correctly). However, all the nonfunctional V segments possess a canonic heptamer and nonamer RSS and thus can potentially rearrange normally with J segments (14). The 60 J segments span a region of 60 kb (15), 17 are considered as pseudogenes, but their status in terms of rearrangement are not clearly elucidated, except J60, 59, 51, 29, and 25 which do not rearrange (16).

Although gene rearrangement is only the first step involved in the establishment of the TCR repertoire it is frequently the less considered step in the estimation of total potential TCR diversity. In theory, any given V segment can rearrange with any D and/or J segment, giving a simple combinatory formula: "rearrangement number = x(V) × y(D) × z(J)," where x, y, and z correspond to the numbers of each segment present on the locus. Several studies have launched a debate as to whether combination between V, (D), and J gene segments is random or not. Some data, point to a biased expressed TCRα chain repertoire with a preferential rearrangement of the most J-proximal V segments to the closest J segments and of the most J-distal V segments to the furthest J segments (17–20). In addition, use of both V and J gene segments appears developmentally regulated (17, 21). Thus, the progressive and coordinated utilization of the TCR-V and -J genes points to a rule resembling a modified version of the bidirectional, coordinated nibbling model proposed for Ig genes (22). However, a locus-wide analysis of TCRA rearrangements in human and mouse αβ T cell clones reported a loose correlation between the 5′ or 3′ position of the V and J segments used in a given rearrangement, leading to the conclusion that the erratic use of V genes appears inconsistent with the bidirectional and coordinated nibbling model (23). Finally to date, available informations encompass essentially either analysis at the transcriptional level or gene analysis for only a few V families, thus precluding a general synthetic overview of gene rearrangements.

In this report, to eliminate the biases due to transcriptional regulation and the effect of positive selection on the emergence of T cell clones, we have used a sensitive multiplex PCR assay at the genomic DNA level. This allows a systematic, qualitative screening of the rearrangements effected by 22 out of the 29 V families in a whole thymus. We have studied samples from BALB/c mice at a number of time points throughout ontogeny. In addition, for certain V-J rearrangements, these rearrangement events were analyzed quantitatively, again at the DNA level. Taken together, our findings indicate that the number of V-J combinations is lower than that predicted by a random rearrangement model and that the combinatory diversity of the TCRα chain is significantly skewed during mouse T cell development.

Materials and Methods

Nomenclature. Nomenclatures for V genes and for J segments are according to ImMunoGeneTics (IMGT) database (http://imgt.cnusc.fr). NCBI accession nos.: for V region, AE008683-AE008686, for J region, M64239.

Mouse. BALB/c mice were purchased from IFFA CREDO. RAG-$2^{-/-}$ (24) were raised in our SPF animal facility. Fetal thymi were obtained from timed pregnancies, where fetal day 1 corresponds to the day of detection of a vaginal plug. Thymic lobes from embryonic or neonatal mice were pooled and mechanically dissociated in PBS before DNA extraction.

Fetal Thymic Organ Culture. BALB/c fetal thymi were isolated on day 16 of gestation and placed on 0.8-μm pore size membranes (Pall Corporation) floating on complete IMDM medium in 24-well plates, as described (25). After 6 d in culture, DNA was extracted from thymic lobes.

Multiplex PCR Assay. Genomic DNA was extracted and amplified as described (9). Multiplex PCR were performed as described (16). Briefly, using an upstream primer specific for a given Vx family and a downstream primer specific for a given Jy segment, the multiplex PCR assay allows the detection of a Vx to Jy rearrangement as well as that of Vx genes to a limited set of 5′ J segments spanning from Jy to Jy-4 position. The multiplex PCR strategy could potentially amplify DNA fragments corresponding to rearrangement of a Vx+1 gene 3′ to that targeted by the Vx specific oligonucleotide. Actually, we did not detect amplicon corresponding to a fragment containing such rearrangement, ex-

cept for one band at 3594pb obtained with V3x primer which is compatible with a rearrangement of V4 gene located downstream to the targeted V3x gene. Amplifications were performed with 1.3 U/reaction of Expend High Fidelity PCR system (Roche Diagnostics). The cycling conditions were 5 min at 94°C, 26 cycles of 1 min at 94°C, 1 min at 58°C, 6 min at 72°C, and one cycle of 10 min at 72°C. In these assay conditions, maximum amplicon size was ~5 kb. Normalization of the quantity of DNA in each reaction was determined by amplification of either a Cα gene exon or the p53 gene in the same PCR run. Negative PCR controls included DNA isolated from RAG-2$^{-/-}$ thymi, in which no rearrangements take place. Oligonucleotides specific for V, J segments, and Cα segment were as described (9, 16, 17) with additional primers (Table I).

PCR products were separated on 1.5% agarose gels and Southern blotted. Higher bands were separated on 0.6% agarose gel during 24 h. To confirm the specificity of the hybridization signal, the same membrane was successively hybridized with V- and J-specific probes. Amplified fragments were considered as specific products of rearranged gene segments when they migrated at the expected position and were hybridized with both appropriate V- and J-probes. Autoradiograms were exposed and quantitative analysis performed on a Personal Molecular Imager FX (Bio-Rad Laboratories) using Quantity One 4.2.1 software (Bio-Rad Laboratories).

Quantitative PCR. PCR were performed on a Light Cycler™ (Roche Diagnostics) using the FastStart kit, 1 U/reaction. 25 ng matrix DNA was diluted in 25 ng of salmon sperm DNA. The cycling conditions were as follows: 94°C 10 min, 40 cycles (94°C 15 s, 60°C 15 s, 72°C 10 s). Melting curves of PCR products were determined according to the manufacturer's instructions. With a PCR elongation time of 10 s, only the rearrangement to the J segment closest to the downstream J primer is amplified. The specificity of the unique amplification product was determined by melting curve analysis and by migration on agarose gels followed by Southern blotting.

Table I.

Primer name	Sequence (5′–3′)	S/A
AJ56do	TCAAAACGTACCTGGTATAACACTCAGAAC	A
AJ40	TCTTTCTGCTTAACCTGTCCCTCATG	A
AJ33do	TTAGCTTGGTCCCAGAGCCCC	A
AJ33	CATGCATTATTCAGCCAGTGCCTTCT	A
AJ16	CTTGCTTCCCGTGATGTCTGGATGA	A
AJ9do	ACCGAAAGCCACAGGTAACTCCTTATCTCC	A
AJ2	TACCGGGTTGCAAATGGTGCCACTT	A
X6.5 (P53)	ACAGCGTGGTGGTACCTTAT	S
X7 (P53)	CACATGTACTTGTAGTGGATGG	A
Probe name		
AJ40p	TCTTGCTTGTACTACTTACGT	
AJ16p	ACCCCACTAACATGTCTAAAAG	
AJ9p2	GTAATTTAAATCAAGTTTCTCATTGCACTC	
AJ2p2	GGGGTAACTCACCAGATATTACCGTCACT	

S, primer in transcription orientation; A, primer in reverse of transcription orientation.

Results

Global Analysis of V-J Rearrangements in the Adult Thymus. To determine the number of V-J recombinations, we performed a systematic analysis of the genomic rearrangements of 22 V families with a defined set of J segments. We used a multiplex PCR technique allowing one-tube amplification of DNA rearrangements of a given V family with four to five juxtaposed J segments contained within a targeted region. We first focused on two J segments, J56 located 5′ of the J region and close to the V segments and J27 in the center of the J region.

Detailed analysis of the rearrangement patterns of adult thymocytes reveals that V families can be divided into four groups (Fig. 1, a and b). The group I which is the major group includes most of the V families located in the middle part of the V locus. These families comprise 2 to 10 members each, rearranging to approximately the same extent with both 5′ and central J segments. The group II encompasses V12 and V19, located on the 5′ distal end of the V locus. These segments rearrange with the central J27 to J30 segments only. The last two groups contain the V families most proximal to the J region. Group III is composed of V16 and V20, which rearrange mainly with J56 to J61 segments but also rearrange weakly with J segments from the central region. Group IV is restricted to V6 and DV101, 102, and 105, families which are only found rearranged with the most proximal J56 to J61 segments. The same pattern of rearrangement was detected in five independent preparations of thymocytes from individual adult mice, indicating that this global rearrangement pattern is identical from one individual thymus to another.

Comparison of V-J Rearrangements Between Fetal and Adult Mice. To determine if this global rearrangement pattern is a property of the adult thymus or whether it is already established during fetal life, the same experiment was performed with DNA extracted from fetal thymi at 18 d of gestation (F18; Fig. 1, c and d). A comparison of adult and F18 rearrangement patterns reveals that: (a) rearrangement signals from fetal samples are much weaker than those in adult samples; (b) while rearrangements amplified with the J56 primer are readily detectable in the F18 sample (Fig. 1 c), no rearrangements are detected with the J27 primer (Fig. 1 d); (c) the V families found not to rearrange with J56 to 61 segments in the adult thymus (V12, 19) are also found not to do so in fetal samples.

According to Cα gene detection F18 sample was more concentrated than postnatal day 28 (D28) sample (Fig. 1 e). Prolonged exposure of hybridized filters did not reveal additional rearrangement bands in Fig. 1, c and d (unpublished data). The difference in intensity of signal between F18 and young adult D28 samples therefore reflects differences in rearrangement levels.

The J Region Becomes Gradually Accessible. The preceding results indicate differential use of the J segments in fetal and adult mice. To extend our observations we analyzed V2-J rearrangements during the early stages of thymic development in more detail. The six V2 family members,

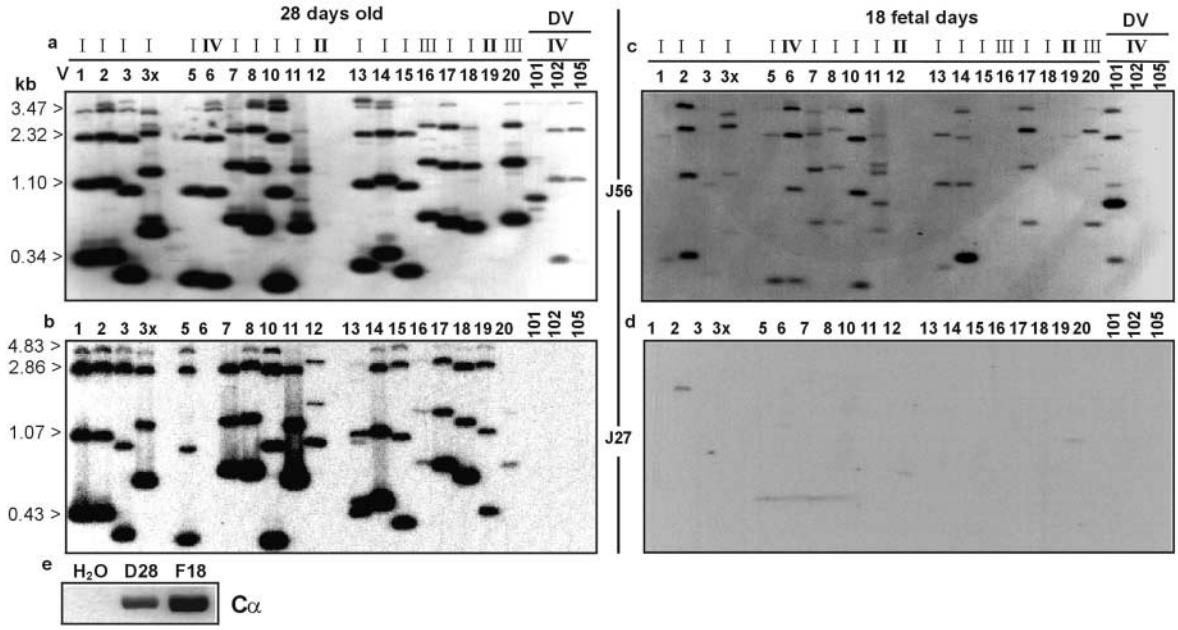


Figure 1. Analysis at F18 and D28 of rearrangements of 22 V families with J56 and J27 primers. Multiplex PCR reactions were performed on genomic DNA using individual V primers in conjunction with primers downstream of J56 (a and c) and J27 (b and d). Products were separated by gel electrophoresis, Southern blotted, and probed with radioactively labeled oligonucleotides specific for either J56 or J27. Each band corresponds to a rearrangement event as determined by distance migration. (e) For each PCR reaction, amplification of the Cα exons served as a positive control. The quantity of DNA in F18 samples was increased compared with D28 in order to maximize the possibility of amplifying rearranged fragments.

which are dispersed over the V region, were amplified with a consensus V2 oligonucleotide in a multiplex assay using nine different J segment primers spread over the J region, as shown in Fig. 2. Our assay allows us to refine the analysis of the number of potential J segments used in V-J rearrangements. Until now, the status of J segments in terms of pseudogenes, which can be defined by an inability to rearrange and/or to be translated as a functional protein, has not been completely elucidated. As seen in Fig. 2, 34 different J segments were found rearranged with V2. A longer migration of the gel allowed us to detect a total of 49 J segments successfully rearranged (unpublished data). Among the 11 remaining nonrearranging J segments, seven have previously been described as incapable of rearrangement: J36 and J1 (IMGT http://imgt.cnusc.fr), and J60, J59, J51, J29, and J25 (16). Our study confirms these results and identifies four additional J segments unable to rearrange with members of the V2 family, namely J20, J19, J14, and J3. On the other hand, J61 which has been denoted as a nonrearranging pseudogene is detected in rearrangements in BALB/c mice.

In our experimental conditions, rearrangements are detected from the 18th day of gestation with J segments in the 5′ quarter of the J gene cluster (Fig. 3). Eight out of a possible eight V2-J rearrangements are readily detected in this region, a further two rearrangements are detectable with J40 and J34 while no rearrangements are detected beyond J34. From fetal day 19, the J locus opens up toward the 3′ end of the V region and rearrangements involving J17 become detectable. From day 20 of gestation, rearrangements are detected up to J2 indicating that the whole J region is accessible for recombination. From this time point, band intensities for the 3′ J segment rearrangements increases

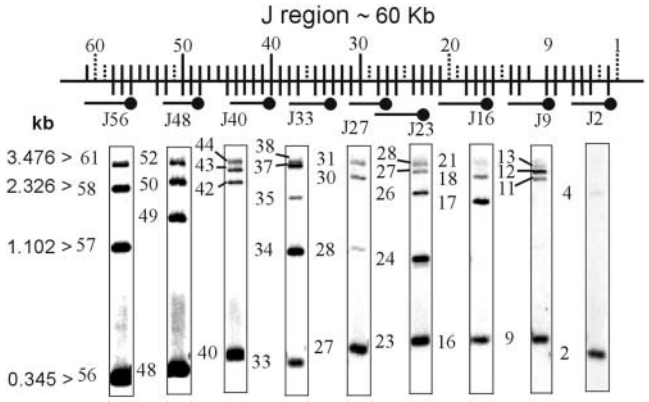


Figure 2. Multiplex analysis of J rearrangements for V2 genes. Multiplex PCR products were Southern blotted and probed with a pool of radioactively labeled primers specific for the nine J segments. Multiplex PCR analysis was performed as described in Fig. 1. Molecular weight indicators are given to the left of the figure. Individual lanes correspond to individual PCR reactions. Each band corresponds to a rearrangement with the J segment indicated to the left of the lane. The 9 J segments chosen are spread along the J region as indicated. Representation of J region (not to scale): dots indicate the position of J primers and probes, the corresponding lines indicate the range of J segments detectable in each reaction lane. The J region is composed of 43 functional J segments on a total of 60 segments. Dashed lines indicate 6 untranscribed J pseudogenes, J55, 54, 47, 46, 11, 7 according to IMGT database (http://imgt.cnusc.fr); cross lines indicate 11 unrearranged J pseudogenes according to IMGT and correlated with our data, J60, 59, 51, 36, 29, 25, 20, 19, 14, 3, (J1 according to IMGT). J10 has never been described in the mouse.

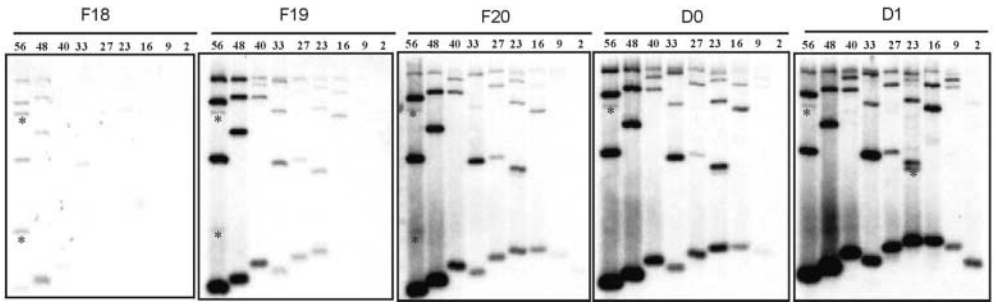


Figure 3. Analysis of rearrangements of the V2 family with J segments during ontogeny. The 9 J segments studied are indicated for each lane. Rearrangements were analyzed by multiplex PCR on DNA extracted from F18, F19, F20, D0 and D1 BALB/c thymi. Products were revealed using an internal V2 probe. Asterisks indicate nonspecific products, as determined by distance migration.

while those of the 5′ J segments remain similar up to 4 wk after birth. However, it is notable that band intensities for rearrangements involving V2 and 3′ distal J segments are lower than those of rearrangements with V-proximal J segments at all time points analyzed. Taken together, our results show a programmed opening of the TCRAD locus in the J region, moving from 5′ J segments, close to the V region, toward 3′ J segments closer to the Cα gene, as has been demonstrated previously in studies at the transcript level (16). This demonstrates that the progressive J usage, as observed by analysis of mRNA is regulated at the level of gene rearrangements.

The V Region Becomes Gradually Accessible. We have previously demonstrated at the transcript level that, in addition to the opening of the J region of the TCRAD locus, there is a similar opening of the V region (17). We further investigated whether V localization is a factor that dictates the rearrangement to particular J segments. We availed of the multiplex technique to study the families identified in Fig. 1 that showed differential rearrangement to J56 or J27 genes. Thus, we compared in adult and newborn thymus two V families situated proximal to the J cluster, V6 (340 kb from Cα coding region) and V20 (380 kb) with two V families distal to the J cluster, V12 (1,420 kb) and V19 (1,580 kb). As expected from the results in Fig. 1, V6 and V20 rearrangements are mainly detected with proximal, whereas V12 and V19 rearrange with a more distal group of J segments (Fig. 4 a). Nevertheless, there is a notable difference in the range of utilization of proximal J segments between V6 and V20. In five different experiments, we detect V6 rearranged to J segments located between J61-J33 and V20 rearranged to J segments located between J6–J23. Taking into account that between the J61 gene and either J33 or J23, 24 or 32 J segments contain a classical RSS, we calculate that V6 and V20 can be recombined with 49 and 65% of the functional J segments, respectively.

Conversely, V12 and V19 show similar patterns of rearrangement: considering that the rearrangements of both these variable genes are observed with a set of J segments from J49 to J16, this suggests that these variable genes can rearrange with ~67% of the J locus.

Furthermore, to test whether the V segments located in the center of the region also use a coordinated set of J segments, we performed the same analysis with V2S2 (1,140 kb from Cα) and V2S7 (490 kb) using specific primers for each gene (Fig. 4 b). As expected, V2S2 rearrangements are centered on J27 and give a very similar pattern to that of V12 and V19. Conversely, V2S7 preferentially uses J48 (Fig. 4 b). To confirm these results, similar analyses using the four V7 family members, whose genes are located between V2S2 and 2S7 segments, show that the V7 genes preferentially use J elements in the middle of J locus (unpublished data).

In addition, the V families closest to the J region (V20 and V6) are detected in their rearranged form as early as F18 whereas those furthest from the J cluster (V19 and V12) only become detectable from day 1 after birth (Fig. 5, and unpublished data). These data are in agreement with the idea that particular V families show preferential rearrangement to certain J segments, both in adult and in newborn thymi, indicating that there is a stable mechanism throughout ontogeny.

V-J Rearrangement Patterns Are Determined in the Thymus Before Birth. Due to the spatial and temporal nature of the V-J rearrangements described above, later-appearing V12- and V19-J rearrangements could be the result of a new wave of T cell progenitors entering the thymus around the time of birth. To test this hypothesis, we investigated the rearrangement of various V families to the J region in DNA from day 16 fetal BALB/c thymic lobes maintained in fetal thymic organ culture (FTOC) for 6 d. The timing of isolation of fetal thymic lobes allowed us to study cells derived from the first wave of precursors entering the thymus. Thymic lobes were cultivated for 6 d in vitro, corresponding in age to those of 1-d-old mice. Very similar V-J rearrangement patterns were observed with DNA originating either from FTOC or from thymi taken from 1-d-old mice (Fig. 4 a). Thus, in DNA extracted from FTOC, V2 genes rearranged to the whole J region, V20 and V6 mainly rearranged to proximal J segments, whereas V12 and V19 were found rearranged to more distal J segments, results similar to those obtained with DNA from newborn and adult mice. These results show that the rules of V-J rearrangement are conserved in FTOC, suggesting that the relatively late V12- and V19-J rearrangements are not induced solely in the progeny of a postnatal wave of thymic precursors. In addition, these data underline the presence of an intrathymic program controlling the opening of the TCRAD locus. This program is independent of signals coming from other organs and of precursors entering the thymus after fetal day 16 (F16).

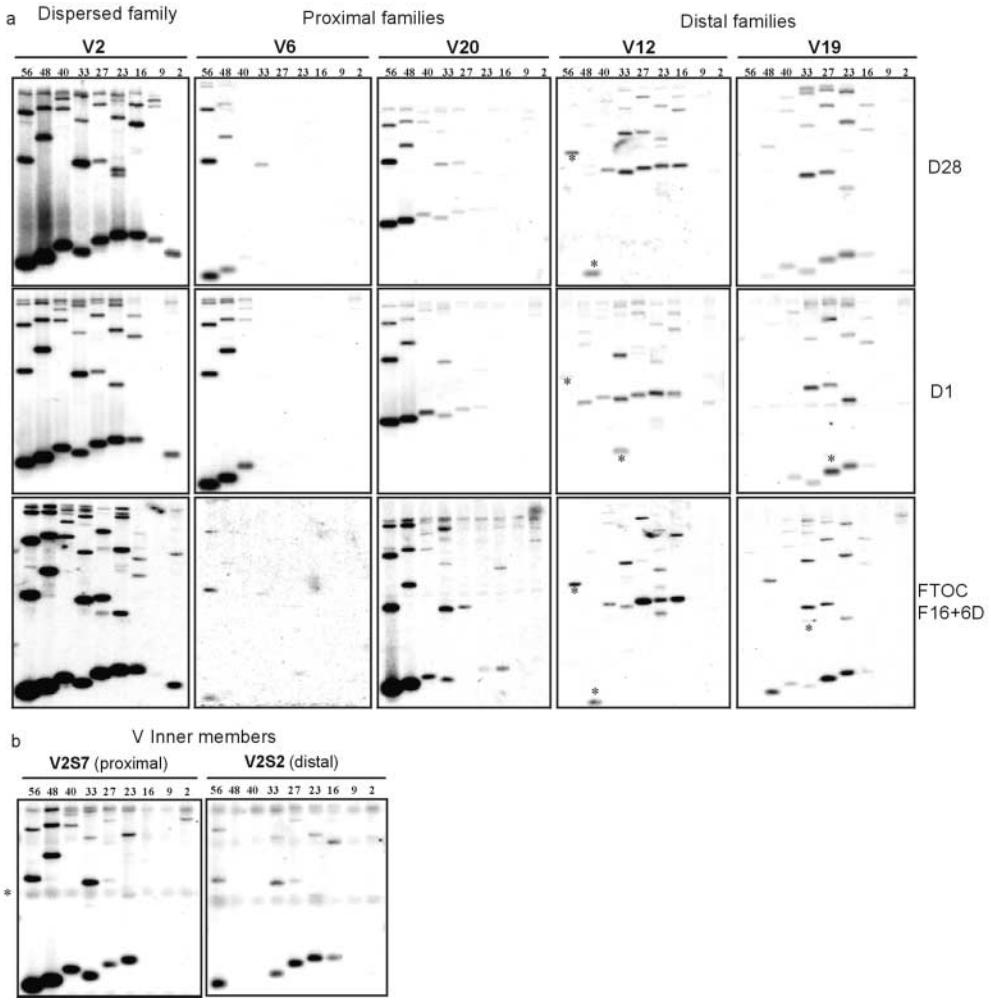


Figure 4. Kinetics of V2, V6, V20, V12, and V19 rearrangements. (a) Rearrangements were analyzed by multiplex PCR on DNA extracted from thymi of 28-d-old, day 1 BALB/c mice, and FTOC from F16 after 6 d of culture. Specific V family primers are indicated on panels, these were used in conjunction with 9 J segments primers, as indicated on lanes. (b) Comparison of V2S2 versus V2S7 members at D28. PCR products were revealed with a pool of appropriate J-segments probes. Asterisks indicate nonspecific products, as determined by distance migration.

Quantitative Differences in V-J Rearrangement. Taken together, our results show that each V gene only rearranges with a restricted set of J segments with differences in the relative use of particular J genes. To analyze these differences more precisely, we have developed a real-time quantitative PCR assay for genomic DNA matrices. This assay was adapted from our previous study where we quantified either V-Cα or V-Cδ transcripts (9).

Rearrangements were analyzed on DNA from the adult (D28) thymus. Nine different PCR reactions were run in parallel using a single oligonucleotide in the V segment coupled with each one of the nine oligos for the J segments. To check the amplification efficiency of the nine J primers, we used the V2 family which, as shown in Fig. 2, rearranges with J segments located across the whole Jα region. The relative order of detection of the nine PCR products using this technique correlates with the relative abundance of rearrangements in the sample as detected by amplification of rearrangements to individual J gene clusters (Fig. 2). Thus, the first product to be detected corresponds to the most abundant rearrangement. Using the Light Cycler™ technology where the quantity of amplicon and fluorescent signal doubles every cycle if efficiency is 100%, the maximal attainable slope for a PCR curve is −3.3. According to the manufacturer's instructions, this slope value is used as follows: $10^{(-1/-3.3)} = 2$, where 2 corresponds to a 100% increase (or doubling) of amplicon yield per cycle. For the nine PCR reactions of V rearrangements, the slopes of the amplification curves were similar, ranging from −3.7 to −3.75, indicating equivalent reaction efficiencies and thus allowing comparison of the levels of V-J rearrangements in a given set of reactions (Fig. 6). The average yield per reaction was 85–86% (calculated as follows: $10^{(-1/-3.7)} = 1.86$ or 86%), equivalent to a 1.7-fold increase of PCR product per cycle. For the V6 gene, only rearrangements to J56, J48, and J40 were detectable (Fig. 7 a). The quantitative comparison between rearrangements involving any two genes is calculated by multiplying their amplification factor (1.7) by the difference in cycle number at which their PCR products were initially detectable. For each V6-J40 rearrangement event, 6 V6-J48 and 6.7 V6-J56 rearrangements were detected. This demonstrates a significant skewing of the frequency of utilization of J segments by a given V gene and suggests a preferential distribution around J56. Similar experiments were performed with the V19 gene, which uses a different set of J segments.

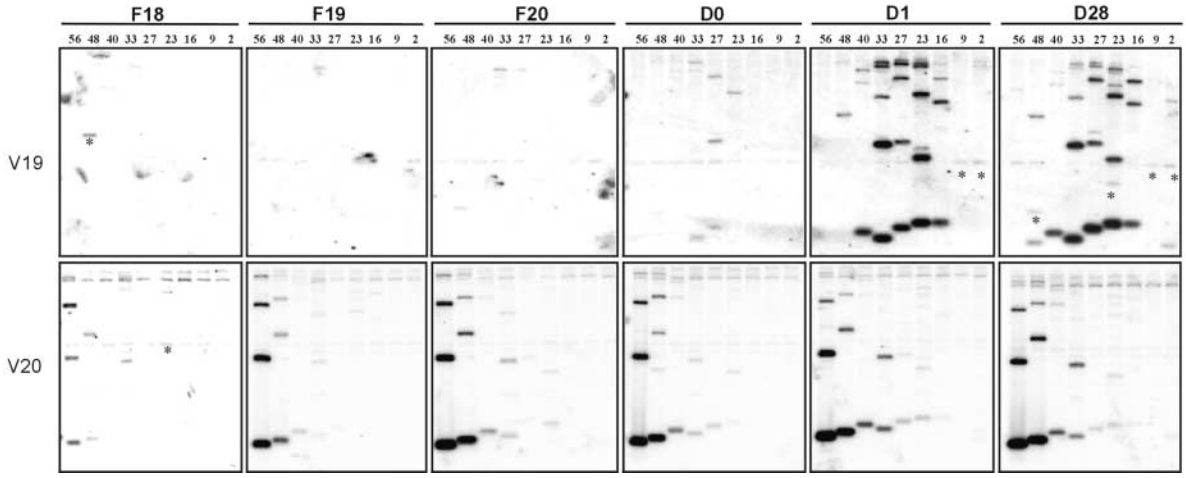


Figure 5. Multiplex analysis of rearrangements for V19 and V20 families during ontogeny. Specific V family primers are indicated on panels, these were used in conjunction with 9 J segment primers, as indicated on lanes. Asterisks indicate nonspecific products, as determined by distance migration.

In this case, we find a symmetric distribution of rearrangements around the J27 segment, suggesting a gaussian pattern of J usage around the central part of the J region. V19-J27 rearrangements were 7.7 times more frequent than rearrangements involving J2. Thus, each V gene rearranges with a given cluster of J segments, but within this cluster, frequencies of rearrangements differ greatly. Taken together, this has a strong influence on the total number of V-J rearrangements contributing to the TCRα repertoire as a whole.

Different V Segments Show Different Frequencies of Rearrangement. The absolute number of DNA molecules corresponding to V6-J56 and V19-J27 rearrangements in a particular sample was calculated. We chose these two rearrangement events as they are the most abundant for the two variable segments studied in the previous section. PCR products corresponding to V6-J56 and V19-J27 were purified and dilutions used as PCR substrate to generate a standard curve alongside test samples. This allows the determination of the number of each rearrangement in the starting population. The DNA contents of test samples were normalized by amplification of p53, as a nonrearranging gene, present at two copies per cell (unpublished data). At F19, 40 V6-J56 rearrangements are detectable in 25 ng DNA, which corresponds to a cell equivalent of 4167 (calculated as follows: 25.10^{-9} g of DNA divided by 6.10^{-12} g of DNA per cell; Fig. 7 b). From F20 to D4, V6-J56 is present at

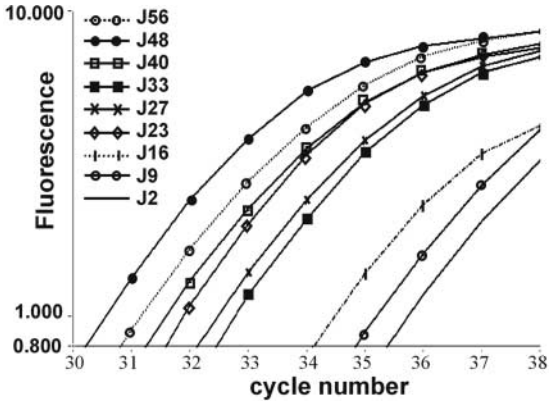


Figure 6. Quantification of V2 rearrangement on 9 J segments, in D28 thymus. V2 family rearrangements were amplified using cDNAU 5′ primer in conjunction with 9 different 3′ primers. The slope of the amplification curves was calculated to determine comparability of data.

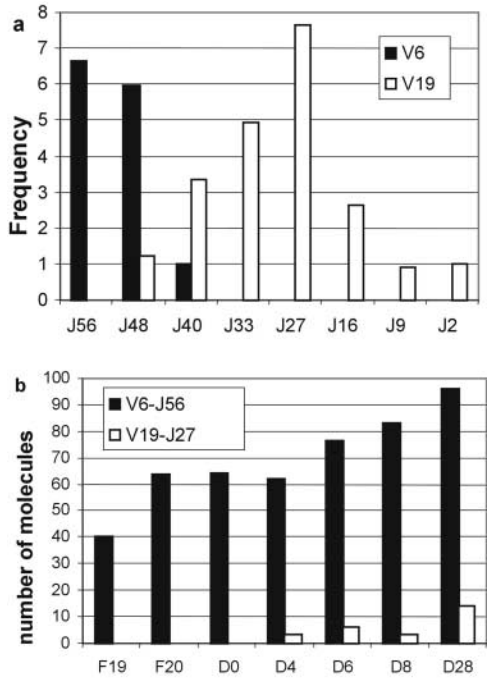


Figure 7. Quantitative genomic rearrangement of V6 and V19 segment. (a) Frequency of utilization of 8 J segments by V6 and V19 in D28 thymus, determined by real time PCR. (b) Quantification of V6-J56 and V19-J27 rearrangement during BALB/c ontogeny. For each rearrangement the exact number of molecules is calculated from the corresponding DNA external standard curve.

~60 molecules before rising from 76 to 96 molecules between D6 and D28. For V19-J27, 4 rearrangements are detectable at D4, rising to 14 molecules at D28. These results show that in adult thymus, 1.15% of the chromosomes in the 4167 cell equivalents have rearranged V6-J56 (calculated as follows: 96 molecules divided by 2×4167 equivalent cells, assuming that both chromosomes are fully rearranged), compared with 0.16% which have rearranged V19-J27. This corresponds to ~7-fold difference between the two variable gene segments, indicating that V6-J56 rearrangement is seven times more frequent than V19-J27 recombination. In addition, the RSS of V19 unlike that of V6, is very similar to the RSS consensus sequence of mouse V α elements (26), indicating that the low level of V19 rearrangement is not due to a defect in its RSS. Taken together, these data point out a significant influence of the distance from the J region on the utilization of V segments.

Discussion

In this report, using multiplex genomic PCR analysis, quantitative and qualitative analysis of representative V-J combinations, we conducted a systematic analysis of the V-J rearrangement profile of normal mouse thymocytes during development. There are four major conclusions from this study. First, the V-J rearrangement pattern is not random either in the fetal or adult thymus thereby imposing severe restrictions on the potential TCRα repertoire. Second, from analysis of FTOC, establishment of V-J rearrangement pattern is intrinsic to the thymus. Third, for individual V genes, rearrangement targets particular sets of J genes and within these particular sets, there is differential usage of individual J genes. Finally, there is quantitatively more rearrangement involving V genes at the 3′ than at the 5′ end of the V locus. Taken together, these results lead us to conclude that accessibility of the TCRA locus is coordinately regulated but that this regulation affects the V and J regions differently.

Previous analyses of V-J rearrangements in mouse thymocytes have used Southern blot analysis of genomic DNA (27). These studies provided initial indications that there was progressive 3′ to 5′ usage of J genes during development. However, due to the difficulty of quantifying a decrease in the hybridization signal from the nonrearranged genomic fragments and weak hybridization signal from rearranged bands, the Southern blot approach is relatively insensitive. Limited PCR analysis of the transcribed repertoire using a restricted set of V and J primers has also been performed. These studies have proposed an ordered and coordinated expression of the TCRAD locus (17–20). To eliminate the biases due to transcriptional regulation, we have used a recently developed multiplex PCR analysis of genomic DNA (16). The added sensitivity of this approach allows us to draw important conclusions regarding the development and diversity of the murine TCRα repertoire. The screening of all V families located closest to the Cα coding region (between −400 and −345.7 kb) indicates that these V segments rearrange predominantly with the most proximal J segments. Reciprocally, the V families situated in the most distal part of V region (between −1584.2 and −1422.2 kb) are preferentially rearranged with the J segments found in the midsection and the most distal parts of the J region. This indicates a programmed opening of the V region from 3′ to 5′, incompatible with the idea of a random repartition of recombination during thymocyte development. Consequently, it is not surprising that during fetal life only the rearrangements between the 3′ V and the 5′ J segments are detected. However, it has been proposed for the regulation of the human TCRAD locus, that the coordinated model concerns solely the V segments located at the extremities of the V region (23). The following grounds could explain the apparent disagreement. The human V rearrangements analyzed were derived from a panel of T cell clones. The rearrangement events were studied by cloning and sequencing of RT-PCR products or from published sequences. In our case, the whole study was done ex vivo at the genomic DNA level and at different stages of thymus development. This approach gives a direct view of rearrangement status and decreases the bias due to thymic selection. The human locus is less complex in term of V elements than its mouse counterpart and therefore the possibility of analyzing coordinated and polarized V segment use is restricted. In spite of the high level of homology between the two loci, some structural constraints may be divergent, leading to the emergence of different rules or factors governing the accessibility of the V and J regions between the human and the mouse loci.

Recently we proposed, using RNA level quantification, that TCRα chains are produced in regulated waves during thymic development (9). These regulated waves could corroborate with either the set of new thymic immigrants during ontogeny or an internal regulation in the TCRAD locus. Our analyses of TCRα rearrangement in FTOC are more in accordance with the second hypothesis, as a similar profile of TCRα repertoire was observed between isolated and in vivo thymocytes. This assumes that the rearrangements of TCRAD locus are controlled within the locus and depend mostly on the strict regulation of the opening and accessibility of chromatin.

As reported in this work, during rearrangement each individual V gene targets specific sets of J genes, leading to the restriction of the potential V repertoire. Thus, among the J set used by a given V segment, the J segments are used with a gaussian distribution centered on about 15 J segments. This differential frequency leads to the emergence of preferential recombinations and consequently to a reduction of TCRα diversity. Thus, before pairing with TCRβ chain and before the selection process during T cell development, the genomic repertoire of TCRα chain is already restricted. As we show here that all the rearrangements between V and J libraries are not equivalent and that their frequencies are diverse, one can assume that the diversity of TCRα chain expression is probably lower than the 1.18×10^4 proposed by Cabaniols et al. (28). Taking into account that a given V gene can rearrange with a maxi-

mum of 67% of J segments, we propose that the V diversity is in the order of $0.67 \times 1.18 \times 10^4$. Thus, assuming a TCR$\beta$ diversity in the order of 6×10^5 (29), the number of potential $\alpha\beta$ combinations should be inferior to 4.7×10^9. It will be interesting now to determine whether the predominant V-J rearrangements have an influence on thymocyte maturation and thereby limit the repertoire of TCR$\alpha\beta$ actually expressed.

We attempted for the first time to make an accurate comparison of the relative rearrangement and transcriptional status of a single V gene (V19) located at the 5′ extremity of the locus. This analysis revealed that in adult mice, whereas the V19 gene is rearranged sevenfold less than V6, nevertheless, as we have recently reported (9), V19 transcripts are twice as abundant as those for V6. Likewise, we detected DV101, 102, and 105 rearrangements to proximal J genes, but again at the transcription level they were barely detected. During fetal life, initial TCRA rearrangements utilize a spatially restricted set of V genes located at the 5′ end of the J region. Taken together, these results indicate that during development, 3′ V region genes become accessible to rearrangement before those in the more distal 5′ end of the locus. This would give an initial 3′ V bias to the expressed V repertoire. As proposed previously (30, 31) use of more distal V segments more than likely occurs as a result of secondary V to J rearrangements. This process would normally occur at the DP stage of thymocyte development and its extent would be limited by cell survival time as well as the duration of RAG gene expression and protein function. Studies recently published (32, 33) would indicate that secondary V-J rearrangements favor utilization of J genes close to those used in the initial rearrangement. In this scenario the combined activity of the Eα enhancer, TEA element and promoter of the initially rearranged V gene, result in a remodeling of V/J chromatin thereby allowing a new set of 3′ J segments to become accessible to secondary recombination.

For the V region, recent data as well as those reported in this study favor a model of progressive tracking down the V locus. The primary rearrangement would mainly involve 3′ V genes either because a substantial portion of the V and J region becomes accessible to RAG activity or because the first V genes are rearranged behind an initial V-D rearrangement (34). For subsequent rearrangements, it is unclear whether the entire remaining V locus now becomes accessible or whether small windows are sequentially opened. We favor the former hypothesis given that: (a) contrary to the J region (35, 36), there is no interallelic coincidence in the chromosomal position of the two rearranged V genes (23); (b) the number of secondary rearrangements is limited by DP thymocyte lifespan; and (c) individual V genes show a gaussian-like utilization of J genes. To account for these observations, it would seem that the V region would be scanned faster than the J region in order for DP thymocytes to rearrange their most distal V genes. This would also explain why 3′ V genes rearrange primarily to 5′ J genes whereas distal 5′ V genes use both central and 3′ J elements (Fig. 8).

For the J locus, Guo et al. (32) have recently proposed a "local service" model to account for the pattern of J gene utilization, suggestive of sequential V-J rearrangements proceeding in small steps to J segments adjacent to the V-J joint to be replaced. This "local service" was in comparison to the term "express service" which they dismissed as a way of describing J gene utilization. However, in order to take into consideration kinetic and topographical aspects of V gene rearrangement, we would favor the "express service" model to describe V gene utilization. For an "express service" model to function there needs to be a large window of gene accessibility in the V locus, a situation more compatible with control by enhancer or enhancer-like elements rather than by V gene promoter activity. Overall, our results are in agreement with a bidirectional and coordinated model of TCRA recombination thereby generating TCRA diversity, yet maximizing the possibility of secondary rearrangements. When applied to situations of altered thymocyte development, the comprehensive quantitative and qualitative approach described herein will be useful for dissecting the complex molecular control involved in TCRAD locus rearrangement.

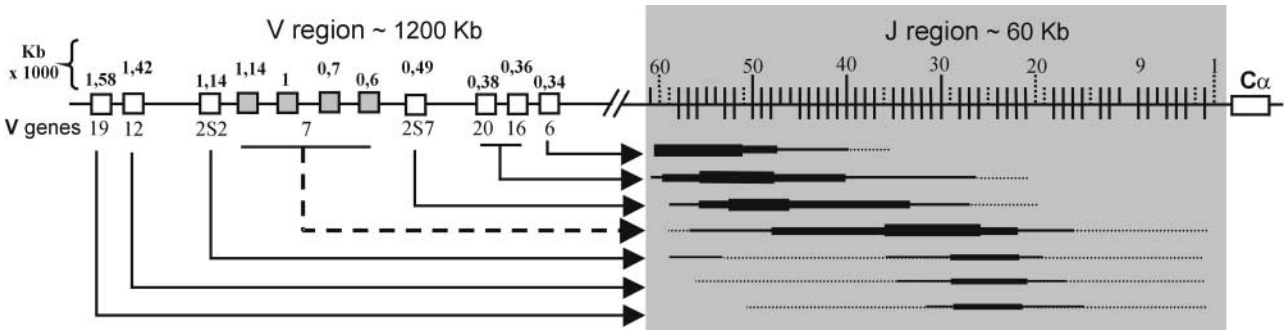


Figure 8. Representation of V-J segment utilization in adult thymus. The V region (not to scale): V member position is indicated by open boxes. V7 family is in gray boxes. Distances (in kb) indicated above V segments are measured from the last base of V gene and first base of Cα exon 1. The thickness of the line under the J region of the locus is an indication of the frequency of J utilization by the different V segments studied. The J region (not to scale) is presented as described in Fig. 2. V20 and V16 show the same distribution of rearrangement. The four V7 members located between V2S2 and 2S7 show an intermediate rearrangement pattern when compared with these two genes.

We thank Eve Borel for technical assistance. We sincerely thank Serge Candéias, Cédric Touvrey, Stéphane Mancini, Bernard Malissen, Jean-Pierre De Villartay, Marc Bonneville, and Thierry-Pascal Baum for helpful discussion and comments.

This work was supported by institutional grants from Institut National de la Santé et de la Recherche Médicale, from Commissariat à l'Energie Atomique (CEA), and from a specific grant "Thématiques Prioritaires de la Région Rhône-Alpes." N. Pasqual is recipient of a fellowship from the CEA.

Submitted: 27 June 2002
Revised: 26 August 2002
Accepted: 15 September 2002

References

1. Clevers, H., B. Alarcon, T. Wileman, and C. Terhorst. 1988. The T cell receptor/CD3 complex: a dynamic protein ensemble. *Annu. Rev. Immunol.* 6:629–662.
2. Gellert, M. 1997. Recent advances in understanding V(D)J recombination. *Adv. Immunol.* 64:39–64.
3. Schatz, D.G., M.A. Oettinger, and M.S. Schlissel. 1992. V(D)J recombination: molecular biology and regulation. *Annu. Rev. Immunol.* 10:359–383.
4. Lieber, M.R. 1992. The mechanism of V(D)J recombination: a balance of diversity, specificity, and stability. *Cell.* 70:873–876.
5. Bogue, M., S. Gilfillan, C. Benoist, and D. Mathis. 1992. Regulation of N-region diversity in antigen receptors through thymocyte differentiation and thymus ontogeny. *Proc. Natl. Acad. Sci. USA.* 89:11011–11015.
6. Sleckman, B.P., C.H. Bassing, C.G. Bardon, A. Okada, B. Khor, J.C. Bories, R. Monroe, and F.W. Alt. 1998. Accessibility control of variable region gene assembly during T-cell development. *Immunol. Rev.* 165:121–130.
7. Capone, M., R.D. Hockett, Jr., and A. Zlotnik. 1998. Kinetics of T cell receptor beta, gamma, and delta rearrangements during adult thymic development: T cell receptor rearrangements are present in CD44(+)CD25(+) Pro-T thymocytes. *Proc. Natl. Acad. Sci. USA.* 95:12522–12527.
8. von Boehmer, H., I. Aifantis, J. Feinberg, O. Lechner, C. Saint-Ruf, U. Walter, J. Buer, and O. Azogui. 1999. Pleiotropic changes controlled by the pre-T-cell receptor. *Curr. Opin. Immunol.* 11:135–142.
9. Gallagher, M., P. Obeid, P.N. Marche, and E. Jouvin-Marche. 2001. Both TCR alpha and TCR delta chain diversity are regulated during thymic ontogeny. *J. Immunol.* 167: 1447–1453.
10. Jouvin-Marche, E., M.G. Morgado, N. Trede, P.N. Marche, D. Couez, I. Hue, C. Gris, M. Malissen, and P.A. Cazenave. 1989. Complexity, polymorphism, and recombination of mouse T-cell receptor alpha gene families. *Immunogenetics.* 30:99–104.
11. Arden, B., S.P. Clark, D. Kabelitz, and T.W. Mak. 1995. Mouse T-cell receptor variable gene segment families. *Immunogenetics.* 42:501–530.
12. Gahery-Segard, H., E. Jouvin-Marche, A. Six, C. Gris-Liebe, M. Malissen, B. Malissen, P.A. Cazenave, and P.N. Marche. 1996. Germline genomic structure of the B10.A mouse Tcra-V2 gene subfamily. *Immunogenetics.* 44:298–305.
13. Cook, G.P., and I.M. Tomlinson. 1995. The human immunoglobulin VH repertoire. *Immunol. Today.* 16:237–242.
14. Sakano, H., K. Huppi, G. Heinrich, and S. Tonegawa. 1979. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature.* 280:288–294.
15. Koop, B.F., and L. Hood. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7:48–53.
16. Mancini, S.J., S.M. Candeias, J.P. Di Santo, P. Ferrier, P.N. Marche, and E. Jouvin-Marche. 2001. TCRA gene rearrangement in immature thymocytes in absence of CD3, pre-TCR, and TCR signaling. *J. Immunol.* 167:4485–4493.
17. Jouvin-Marche, E., C. Aude-Garcia, S. Candeias, E. Borel, S. Hachemi-Rachedi, H. Gahery-Segard, P.A. Cazenave, and P.N. Marche. 1998. Differential chronology of TCRADV2 gene use by alpha and delta chains of the mouse TCR. *Eur. J. Immunol.* 28:818–827.
18. Thompson, S.D., J. Pelkonen, and J.L. Hurwitz. 1990. First T cell receptor alpha gene rearrangements during T cell ontogeny skew to the 5′ region of the J alpha locus. *J. Immunol.* 145:2347–2352.
19. Huang, C., and O. Kanagawa. 2001. Ordered and coordinated rearrangement of the TCR alpha locus: role of secondary rearrangement in thymic selection. *J. Immunol.* 166:2597–2601.
20. Aude-Garcia, C., M. Gallagher, P.N. Marche, and E. Jouvin-Marche. 2001. Preferential ADV-AJ association during recombination in the mouse T-cell receptor alpha/delta locus. *Immunogenetics.* 52:224–230.
21. Rytkonen, M.A., J.L. Hurwitz, S.D. Thompson, and J. Pelkonen. 1996. Restricted onset of T cell receptor alpha gene rearrangement in fetal and neonatal thymocytes. *Eur. J. Immunol.* 26:1892–1896.
22. Wood, C., and S. Tonegawa. 1983. Diversity and joining segments of mouse immunoglobulin heavy chain genes are closely linked and in the same orientation: implications for the joining mechanism. *Proc. Natl. Acad. Sci. USA.* 80:3030–3034.
23. Davodeau, F., M. Difilippantonio, E. Roldan, M. Malissen, J.L. Casanova, C. Couedel, J.F. Morcet, M. Merkenschlager, A. Nussenzweig, M. Bonneville, and B. Malissen. 2001. The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. *EMBO J.* 20:4717–4729.
24. Shinkai, Y., G. Rathbun, K.P. Lam, E.M. Oltz, V. Stewart, M. Mendelsohn, J. Charron, M. Datta, F. Young, A.M. Stall, et al. 1992. RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell.* 68: 855–867.
25. Ceredig, R. 1988. Differentiation potential of 14-day fetal mouse thymocytes in organ culture. Analysis of CD4/CD8-defined single-positive and double-negative cells. *J. Immunol.* 141:355–362.
26. Glusman, G., L. Rowen, I. Lee, C. Boysen, J.C. Roach, A.F. Smit, K. Wang, B.F. Koop, and L. Hood. 2001. Comparative genomics of the human and mouse T cell receptor loci. *Immunity.* 15:337–349.
27. Fondell, J.D., J.P. Marolleau, D. Primi, and K.B. Marcu. 1990. On the mechanism of non-allelically excluded V alpha-J alpha T cell receptor secondary rearrangements in a murine T cell lymphoma. *J. Immunol.* 144:1094–1103.
28. Cabaniols, J.P., N. Fazilleau, A. Casrouge, P. Kourilsky, and J.M. Kanellopoulos. 2001. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J. Exp. Med.* 194:1385–1390.

29. Casrouge, A., E. Beaudoing, S. Dalle, C. Pannetier, J. Kanellopoulos, and P. Kourilsky. 2000. Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes. *J. Immunol.* 164:5782–5787.
30. Wang, F., C.Y. Huang, and O. Kanagawa. 1998. Rapid deletion of rearranged T cell antigen receptor (TCR) Valpha-Jalpha segment by secondary rearrangement in the thymus: role of continuous rearrangement of TCR alpha chain gene and positive selection in the T cell repertoire formation. *Proc. Natl. Acad. Sci. USA.* 95:11834–11839.
31. Petrie, H.T., F. Livak, D.G. Schatz, A. Strasser, I.N. Crispe, and K. Shortman. 1993. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *J. Exp. Med.* 178:615–622.
32. Guo, J., A. Hawwari, H. Li, Z. Sun, S.K. Mahanta, D.R. Littman, M.S. Krangel, and Y.W. He. 2002. Regulation of the TCRalpha repertoire by the survival window of CD4(+)CD8(+) thymocytes. *Nat. Immunol.* 3:469–476.
33. Buch, T., F. Rieux-Laucat, I. Forster, and K. Rajewsky. 2002. Failure of HY-specific thymocytes to escape negative selection by receptor editing. *Immunity.* 16:707–718.
34. de Villartay, J.P., and D.I. Cohen. 1990. Gene regulation within the TCR-alpha/delta locus by specific deletion of the TCR-delta cluster. *Res. Immunol.* 141:618–623.
35. Mauvieux, L., I. Villey, and J.P. de Villartay. 2001. T early alpha (TEA) regulates initial TCRVAJA rearrangements and leads to TCRJA coincidence. *Eur. J. Immunol.* 31:2080–2086.
36. Malissen, M., J. Trucy, E. Jouvin-Marche, P.A. Cazenave, R. Scollay, and B. Malissen. 1992. Regulation of TCR alpha and beta gene allelic exclusion during T-cell development. *Immunol. Today.* 13:315–322.

Annex 2

IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Thierry-Pascal Baum, Nicolas Pasqual, Florence Thuderoz, Vivien Hierle, Denys Chaume, Marie-Paule Lefranc, Evelyne Jouvin-Marche, Patrice-NoEl Marche and Jacques Demongeot. Nucleic Acids Research, 2004, Vol. 32:51-54.

Nucleic Acids Research, 2004, Vol. 32, Database issue ***D51–D54***
DOI: 10.1093/nar/gkh041

IMGT/GeneInfo: enhancing V(D)J recombination database accessibility

Thierry-Pascal Baum*, Nicolas Pasqual[1], Florence Thuderoz, Vivien Hierle[1], Denys Chaume[2], Marie-Paule Lefranc[2], Evelyne Jouvin-Marche[1], Patrice-Noël Marche[1] and Jacques Demongeot

Laboratoire TIMC-IMAG-CNRS UMR 5525, Techniques de l'Imagerie, de la Modélisation et de la Cognition, Université Joseph Fourier, Grenoble 1, Faculté de Médecine, Domaine de la Merci, 38706 La Tronche, France, [1]ICH, Laboratoire d'Immunochimie, CEA-G/DRDC/ICH CEA-Grenoble, INSERM U548 Université Joseph Fourier, Grenoble 1, 17 rue des Martyrs, 38054 Grenoble Cedex 09, France and [2]Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received July 8, 2003; Revised August 30, 2003; Accepted September 18, 2003

ABSTRACT

IMGT/GeneInfo is a user-friendly online information system that provides information on data resulting from the complex mechanisms of immunoglobulin (IG) and T cell receptor (TR) V(D)J recombinations. For the first time, it is possible to visualize all the rearrangement parameters on a single page. IMGT/GeneInfo is part of the international ImMunoGeneTics information system® (IMGT), a high-quality integrated knowledge resource specializing in IG, TR, major histocompatibility complex (MHC), and related proteins of the immune system of human and other vertebrate species. The IMGT/GeneInfo system was developed by the TIMC and ICH laboratories (with the collaboration of LIGM), and is the first example of an external system being incorporated into IMGT. In this paper, we report the first part of this work. IMGT/GeneInfo_TR deals with the human and mouse TRA/TRD and TRB loci of the TR. Data handling and visualization are complementary to the current data and tools in IMGT, and will subsequently allow the modelling of V(D)J gene use, and thus, to predict non-standard recombination profiles which may eventually be found in conditions such as leukaemias or lymphomas. Access to IMGT/GeneInfo is free and can be found at http://imgt.cines.fr/GeneInfo.

INTRODUCTION

The synthesis of the antigen receptors [immunoglobulins (IG) and T cell receptors (TR)] is complex and unique due to DNA molecular rearrangements in multiple loci, located on different chromosomes (1,2). This led to the creation in 1989 of the international ImMunoGeneTics information system® ('IMGT'); a high-quality integrated knowledge resource specializing in IG, TR, major histocompatibility complex (MHC), and related proteins of the immune system of human and other vertebrate species (3). In vertebrates, the four TR loci, TRA, TRB, TRG and TRD, comprise variable (V), diversity (D) (for the TRB and TRD loci) and joining (J) genes, which rearrange in a combinatorial V(D)J way in order to encode, with a constant C gene, the α, β, γ and δ chains, respectively. The TRA/TRD locus organization is even more complex since the TRD locus is nestled within the TRA locus (2,4–6). The loci are shown in more detail in Table 1 (7). The human TRA locus spans 1000 kb and comprises 54 TRAV and 61 TRAJ (2), whereas the mouse TRA locus spans 1550 kb and comprises 98 TRAV and 60 TRAJ (6). Consequently, extensive work will be required to analyse all the possible TRA V-J combinations: 3294 (54×61) in human (2) and 5880 (98×60) in mouse (6). The TRB locus spans 620 kb in human and 700 kb in mouse, and comprises 67 and 35 TRBV genes, respectively, and two TRBD and 14 TRBJ genes [(2), and IMGT Repertoire http://imgt.cines.fr]. Analysis of the TRB loci will require the study of 1876 ($67 \times 2 \times 14$) and 980 ($35 \times 2 \times 14$) different TRB V-D-J combinations, respectively. The IMGT/GeneInfo information system is intended to give user-friendly and intuitive access to V(D)J recombination data in immunology. This information is complementary to that given in the IMGT/GENE-DB database, and the IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView tools (3). It is worth noting that IMGT/GeneInfo, developed by TIMC and ICH (also in collaboration with LIGM) is the first example of an external system being incorporated into IMGT. In this paper, we report the first part of this work: IMGT/GeneInfo_TR, which deals with human and mouse TRA/TRD and TRB loci. The IMGT/GeneInfo information system allows researchers working on VDJ recombination not only to decrease the work time on genomic analysis, but also to avoid the possibility of sequence errors, when V, D and J genes are manually extracted from raw data of up to 1550 kb loci. Results are obtained after a simple two-step process, allowing a practical visualization of all the rearrangement

*To whom correspondence should be addressed at 4 rue Thiers, Cour, 38000 Grenoble, France. Email: tpbaum@imag.fr

Table 1. T cell receptor V(D)J genes in IMGT/GeneInfo

TR V(D)J loci	TRDV	TRAV	TRAJ	TRBV	TRBD	TRBJ
Human						
Total no.	3	54	61	67	2	14
Functional		45	50	47	2	13
ORF		1	8	6	0	1
Pseudo		8	3	14	0	0
Locus size (kb)	TRAD: 1000			TRB: 620		
Sources	TRAD: AE000658–AE000662			TRB: L36092		
Mouse						
Total no.	6	98	60	35	2	14
Functional		79	38	21	2	11
ORF		5	12	1	0	2
Pseudo		14	10	13	0	1
Locus size (kb)	TRAD: 1550			TRB: 700		
Sources	TRAD: AE008683–AE008686			TRB: AE00063–AE00065		

Sources: IMGT/LIGM-DB and GenBank.

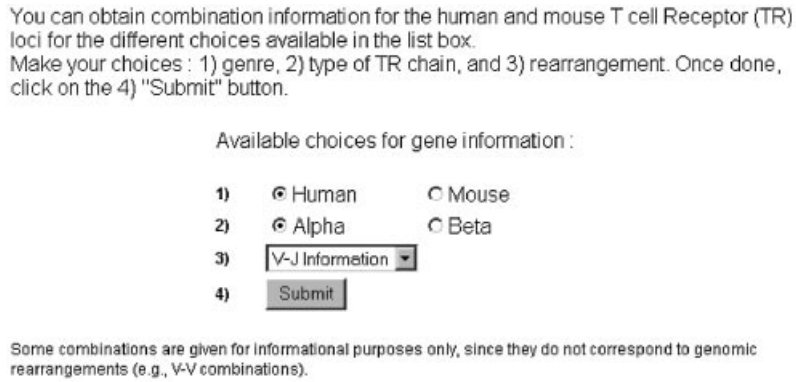


Figure 1. IMGT/GeneInfo query page.

parameters within the same page: gene names, functionality, recombination signal (RS) sequences, locus positions, and sequences of exons and introns.

MATERIALS AND METHODS

IMGT/GeneInfo data extraction

The following references (from GenBank and IMGT/LIGM-DB) were used for data extraction: human (*Homo sapiens*) TRA/TRD (AE000658–AE000662) and TRB (L36092) loci, and mouse (*Mus musculus*) TRA/TRD (AE008683–AE008686) and TRB (AE00063, AE00064, AE00065) loci. Extracted data included the following information for each V, D and J gene: its functionality (functional, pseudogene, ORF), positions of the first and last nucleotide for the gene, V-intron and exon(s) and for the three parts of the recombination signals RS (heptamer, spacer, nonamer). The positions of the V, D and J genes in the TRA/TRD and TRB loci were determined from the first nucleotide of the TRAC and TRBC2 genes, respectively. Data manually extracted from the files were collected for each gene of the six loci. A program automatically extracts nucleotide sequences using the positions of the various elements [gene, V-intron, exon(s), heptamer, spacer, nonamer].

IMGT/GeneInfo query

IMGT/GeneInfo is currently available for the TRA/TRD and TRB loci of human and mouse. The IMGT/GeneInfo query is a two-step process.

Step one: on the first page (Fig. 1), the user selects the species (human or mouse), the locus TRA/TRD (α) or TRB (β) and the gene combinations (V-V, V-J, V-D-J). Some combinations are given for informational purposes only, since they do not correspond to genomic rearrangements (e.g. V-V combinations).

Step two: The second page is generated automatically, and the user then chooses the genes (V, D, J) for which information is required (Fig. 2). Gene choice can be made either according to the gene name [official IMGT nomenclature or previous ones (2,6)], or the relative position of the gene within the locus (e.g. on the TRA locus, position number 1 for the V gene is the most in 5′, and position 1 for the J gene is the most in 3′). All combinations are available, for example, TRAV5 and TRAJ53 (Fig. 2).

IMGT/GeneInfo results

The IMGT/GeneInfo results page is divided into four parts. Reading from top to bottom: Part one is the source from which information was collected (e.g. AE000658 for the human TRA/TRD locus). Part two is an image that corresponds to the

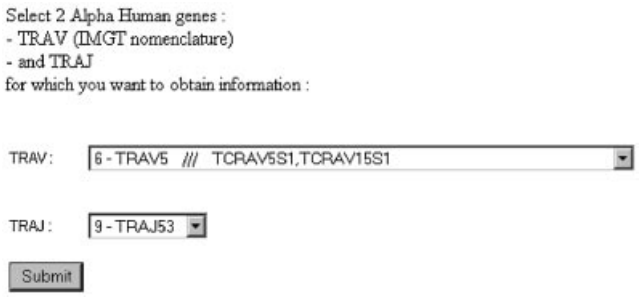


Figure 2. IMGT/GeneInfo gene choice page.

selected combination of genes and that explains visually which gene types are concerned, how the genes and the RS are oriented, and how distances between genes were computed. Part three is a table that contains a summary, for each gene, with the gene name, the functionality (functional, pseudogene, ORF) and the nucleotide sequences for each RS part (heptamer, spacer, nonamer). It also contains the corresponding consensus sequence when it exists; the position relative to TRAC for the TRA/TRD loci and to TRBC2 for the TRB loci; and the genomic distance in base pairs between the genes of the selected combination, in their germline configuration. Part four corresponds to the sequences of the gene and, for a V gene, to its various parts (leader, V-intron, exon 2). These sequences can be selected for copy and paste. A colour code is

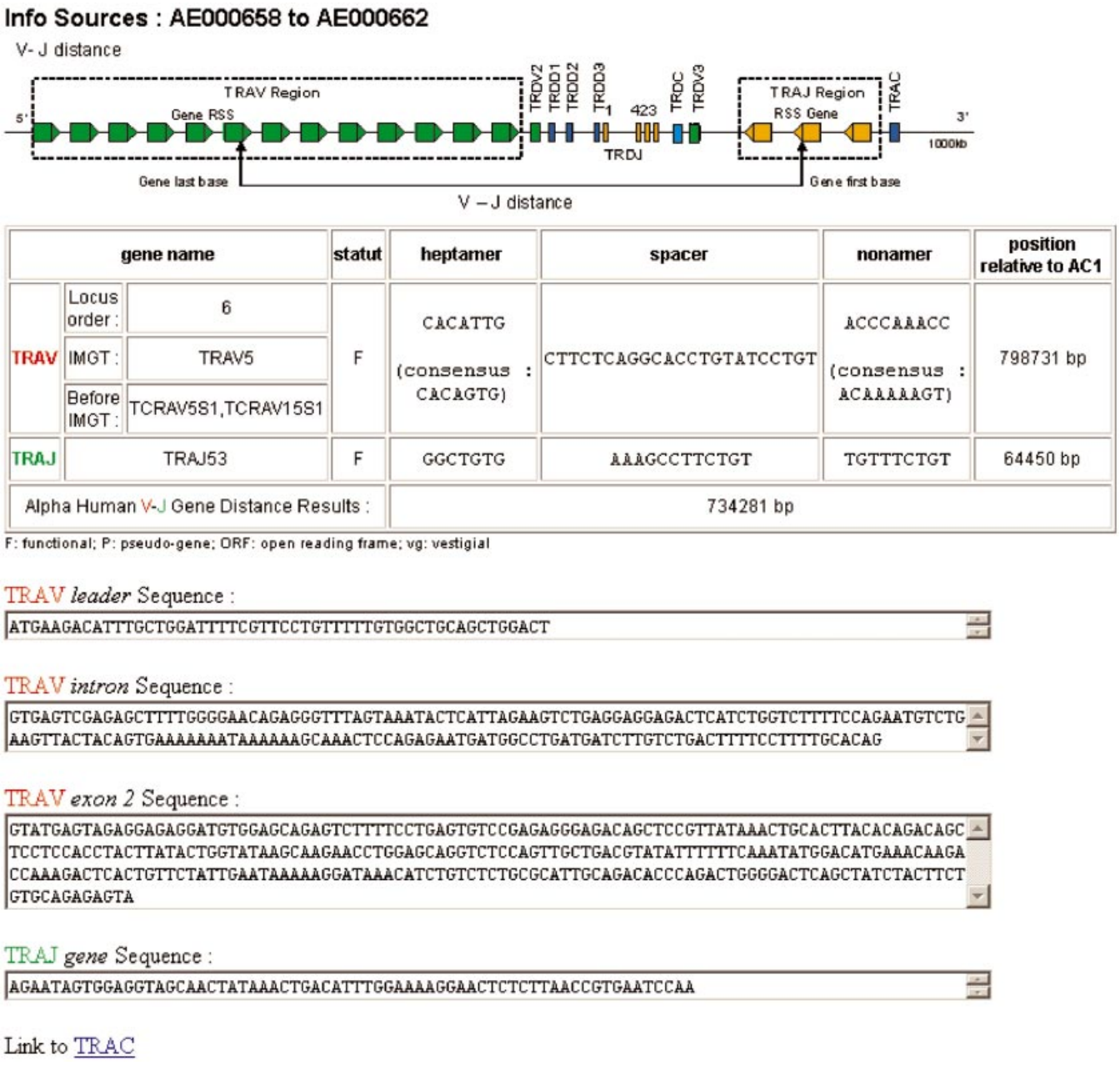


Figure 3. IMGT/GeneInfo results page.

associated with all information originating from the same gene to make it easier to see and remember. A link is provided to the constant gene (e.g. TRAC) from which distances are calculated.

Implementation

IMGT/GeneInfo is deployed in the IMGT information system using Java Servlet technology. The interface uses HTML, JavaScript and CSS.

DISCUSSION AND CONCLUSION

Large genome sequencing allows us to analyse complex loci over few hundred kilobases and to accurately determine their regulation mechanisms. However, raw data utilization in all genetic fields is difficult, and needs a substantial background expertise. This complexity is greatly increased in the IG and TR loci, because of the potential rearrangements of any given V, D and J gene (5). To date, immunologists working on these loci need to manually copy and paste all the potential combinations from sequence databases. The system presented here is the fruit of a collaboration between three laboratories offering complementary backgrounds in immunology, genomics and biocomputing. The IMGT/GeneInfo system allows researchers who work on V(D)J recombinations to greatly decrease the genomic work time as well as to avoid the possibility of sequence errors, working on loci manually shortened to 1550 kb rather than on large raw data. Only two steps are needed to obtain all rearrangement parameters (i.e. gene names, functionality, gene positions, RS, exon and V-intron sequences). The IMGT/GeneInfo information system facilitates easy data archiving. Moreover, because of its ease of use, we expect that this information system will be used as a teaching tool on V(D)J recombination mechanisms.

CITING IMGT/GENEINFO

Authors who use IMGT/GeneInfo are strongly encouraged to cite this article and the IMGT/GeneInfo home page URL, at http://imgt.cines.fr/GeneInfo.

ACCESS AND CONTACT

IMGT/GeneInfo home page: http://imgt.cines.fr/GeneInfo
IMGT/GeneInfo Contact: tpbaum@imag.fr
IMGT home page: http://imgt.cines.fr
IMGT contact: lefranc@ligm.igh.cnrs.fr
TIMC contact: tpbaum@imag.fr
ICH contact: patrice.marche@cea.fr
LIGM contact: lefranc@ligm.igh.cnrs.fr

ACKNOWLEDGEMENTS

We would like to thank Matthew U'Ren-Gerente for his help editing in English. IMGT/GeneInfo is funded by institutional grants from the Institut National de la Recherche Médicale (INSERM), the Commissariat à l'Energie Atomique (CEA) and a specific grant from 'Thématiques Prioritaires de la Région Rhône-Alpes'. The IMGT is funded by the EU 5th PCRDT (QLG2-2000-01287) programme, the Centre National de la Recherche Scientifique (CNRS), and the Ministère de la Recherche et de l'Education Nationale.

REFERENCES

1. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*. Academic Press, London, UK, 458 pp.
2. Lefranc,M.-P. and Lefranc,G. (2001) *The T Cell Receptor FactsBook*. Academic Press, London, UK, 398 pp.
3. Lefranc,M.-P. (2003) IMGT, the international ImMunoGeneTics database®, http://imgt.cines.fr. *Nucleic Acids Res.*, **31**, 307–310.
4. Glusman,G., Rowen,L., Lee,I., Boysen,C., Roach,J.C., Smit,A.F., Wang,K., Koop,B.F. and Hood,L. (2001) Comparative genomics of the human and mouse T cell receptor loci. *Immunity*, **15**, 337–349.
5. Pasqual,N., Gallagher,M., Aude-Garcia,C., Loiodice,M., Thuderoz,F., Demongeot,J., Ceredig,R., Marche,P. and Jouvin-Marche,E. (2002) Quantitative and qualitative changes in V–J α rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor α chain repertoire. *J. Exp. Med.*, **196**, 1163–1173.
6. Bosc,N. and Lefranc,M.-P. (2003) The mouse (*Mus musculus*) T cell receptor α (TRA) and δ (TRD) variable genes. *Dev. Comp. Immunol.*, **27**, 465–497.
7. Gallagher,M., Obeïd,P., Marche,P.N. and Jouvin-Marche,E. (2001) Both TCRα and TCRδ chain diversity are regulated during thymic ontogeny. *J. Immunol.*, **167**, 1447–1453.

Annex 3

Numerical Modelling Of The V-J Combinations Of The T Cell Receptor TRA/TRD Locus.
Thuderoz F, Simonet MA, Hansen O, Pasqual N, Dariz A, Baum TP, Hierle V, Demongeot J, Marche PN, Jouvin-Marche E. PLoS Computational Biology 2010 Vol 6. e1000682. 1-12.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Numerical Modelling Of The V-J Combinations Of The T Cell Receptor TRA/TRD Locus

Florence Thuderoz[1☯], Maria-Ana Simonet[1,2☯], Olivier Hansen[1☯], Nicolas Pasqual[2¤☯], Aurélie Dariz[2,3], Thierry Pascal Baum[1], Vivien Hierle[2], Jacques Demongeot[1*¶], Patrice Noël Marche[2,3*¶], Evelyne Jouvin-Marche[2,3¶]

1 CNRS, Laboratoire TIMC-IMAG, UMR 5525, Grenoble, France, **2** INSERM, Institut Albert Bonniot, Grenoble, France, **3** Université Joseph Fourier-Grenoble I, Faculté de Médecine, Grenoble, France

Abstract

T-Cell antigen Receptor (TR) repertoire is generated through rearrangements of V and J genes encoding α and β chains. The quantification and frequency for every V-J combination during ontogeny and development of the immune system remain to be precisely established. We have addressed this issue by building a model able to account for Vα-Jα gene rearrangements during thymus development of mice. So we developed a numerical model on the whole TRA/TRD locus, based on experimental data, to estimate how Vα and Jα genes become accessible to rearrangements. The progressive opening of the locus to V-J gene recombinations is modeled through windows of accessibility of different sizes and with different speeds of progression. Furthermore, the possibility of successive secondary V-J rearrangements was included in the modelling. The model points out some unbalanced V-J associations resulting from a preferential access to gene rearrangements and from a non-uniform partition of the accessibility of the J genes, depending on their location in the locus. The model shows that 3 to 4 successive rearrangements are sufficient to explain the use of all the V and J genes of the locus. Finally, the model provides information on both the kinetics of rearrangements and frequencies of each V-J associations. The model accounts for the essential features of the observed rearrangements on the TRA/TRD locus and may provide a reference for the repertoire of the V-J combinatorial diversity.

Citation: Thuderoz F, Simonet M-A, Hansen O, Pasqual N, Dariz A, et al. (2010) Numerical Modelling Of The V-J Combinations Of The T Cell Receptor TRA/TRD Locus. PLoS Comput Biol 6(2): e1000682. doi:10.1371/journal.pcbi.1000682

Editor: Rob J. De Boer, Utrecht University, Netherlands

Received September 17, 2008; **Accepted** January 21, 2010; **Published** February 19, 2010

Funding: This work was supported by the institutional Grants from Institut National de la Santé et de la Recherche Médicale (INSERM), from Centre National de la Recherche Scientifique (CNRS), and by the EC Alfa project IPECA. FT was supported by a fellowship from the Agence Nationale de la Recherche et de la Technologie, France. M-AS was supported by a fellowship from Région-Rhone Alpes "Cluster 10". NP was supported by a fellowship from the INSERM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: This work was supported by institutional grants from the Institut National de la Santé et de la Recherche Médicale (INSERM), from the Centre National de la Recherche Scientifique (CNRS), and from a specific grant "Thématiques Prioritaires de la Région Rhône-Alpes". M-AS was recipient of a PhD fellowship from Cluster 10 of the Région Rhône-Alpes; NP was recipient of a fellowship from the INSERM and CEA. FT was recipient of a fellowship from the ANRT (Agence Nationale de la Recherche et de la Technologie) and JD has been supported by the EC Alfa project IPECA. The authors have no conflicting financial interest.

* E-mail: patrice.marche@ujf-grenoble.fr (PNM); Jacques.Demongeot@imag.fr (JD)

¤ Current address: ImmunID, Commissariat à l'Energie Atomique, Grenoble, France

☯ These authors equally contributed to this work.

¶ These authors equally contributed to this work.

Introduction

Functional antigen receptors expressed by T lymphocytes (TR) are generated during ontogeny by somatic recombination of gene segments coding for the variable (V), the joining (J), and the constant (C) segments [1]. The recombination mechanism is largely dependent on both the accessibility of the loci and the RAG enzymatic complex [2–5]. The murine TRA/TRD locus is composite, encoding TR α and δ chains and encompassed of more than 100 functional V genes [6]. In theory, each of the V genes may target one of the 49 functional J genes. The use of V and J genes during the process of recombination has been widely debated, and the studies support the consensus that V-J combinations are not random, with a use of J segments starting at the 5′ end (proximal to the V segments) and proceeding to the 3′ end [7–13]. The accessibility of the J region is controlled by the TR α enhancer (Eα), located at the 3′ end of the C gene [14] and by two promoters: i) T early α (TEA), located at the 5′ end of the Jα region and ii) J49 located 15 Kb downstream of TEA. Both of the promoters are activated by Eα [4,5,15]. Eα controls all the V to J associations whereas the two promoters are required for the rearrangements of the J genes situated at the 5′ end of the Jα region. However, the analyses of TEA-deleted alleles and those of blockade of TEA transcription showed significant alterations in J use and support the hypothesis that the TEA promoter can regulate both positively the promoters located in the first 12 Kb of J genes and negatively the downstream promoters [4,15–17].

A particularity of the TRA locus is an absence of allelic exclusion [18] and its ability to undergo multiple cycles of secondary rearrangements [19,20]. The process of successive rearrangements is stopped by either positive selection, which downregulates recombinase expression [21] or by cell death.

Author Summary

Lymphocytes of the immune system ensure the body defense by the expression of receptors which are specific of targets, termed antigens. Each lymphocyte, deriving from the same original clone, expresses the same unique receptor. To achieve the production of receptors covering the wide variety of antigens, lymphocytes use a specialized genetic mechanism consisting of gene rearrangements. For instance, the genes encoding the receptor of the alpha chain of the T lymphocyte receptor (TRA) spread over a 1500 Kb genetic region which includes around 100 V genes, 60 J genes, and a single C gene. To constitute a functional alpha chain, one of the V and one of the J genes rearrange together to form a single exon. The precise definition of these V-J combinations is essential to understand the repertoire of TRA. We have developed a numerical model simulating all of the V-J combinations of TRA, fitting the available experimental observations obtained from the analysis of TRA in T lymphocytes of the thymus and the blood. Our model gives new insights on the rules controlling the use of V and J genes in providing a dynamic estimation of the total V-J combinations.

Therefore, the impact of secondary rearrangements on the TRα gene assembly regulation remains to be defined.

Regarding the V and J gene use, it is suggested that the first V-J association targets the secondary one into a set of J segments located near the J segment involved in the primary rearrangement [5,16]. The rules governing the use of the V genes have not been clearly elucidated. Nevertheless, observations converge to a consensus: the use of V segments would progress from proximal V genes, located near the J region, towards the V genes located in the distal region [9,10]. At this point in time, the mechanism involved in the control of accessibility of V genes remains to debate [19].

The current state of the technology permits the analysis of some V-J combinations, essentially those at the extremities of the locus but still fails to establish a complete estimation of the V-J combinations. The main obstacle comes from the fact that some V genes are duplicated in similar copies in the V region central part, making problematic their unambiguous identification by molecular methods [22].

Consequently, numerical modelling of the V-J recombination process may offer valuable support to overcome the difficulty for accessing to a global view of TRA repertoire. For instance, if the J genes are chosen in a sequential way in the model, their use results unimodal, whereas it is known from experimental data that TRA/TRD locus displays two Hot Spots of recombination [2–5]. This discrepancy led us to build a mathematical model, parameterized from experimental data, on all V and J genes, including those in distal, proximal, and central positions. Confrontation between the data obtained from experiments and from modelling makes possible an estimation of dynamical parameters, such as the accessibility to rearrangements and the frequencies of the V-J associations, giving a more accurate estimation of the TRA combinatorial diversity.

Results

The goal of building a model representative of the Vα-Jα associations was to reproduce the global biological features of T lymphocyte VαJα rearrangements occurring in the TRA/TRD locus with a software algorithm. This algorithm must be parameterized to find the conditions that reproduce the experimental data. Adequacy between biological and simulated results tells that all the essential aspects of the studied process were included in the model. This modelling approach led us to gather and search for some biological data about the parameters controlling the Vα-Jα rearrangement process in the mouse TRA/TRD locus.

Table 1. TRA/TRD locus characterization.

	V region	J region
Length (Kb)	1300	64
Number of elements	104	60
Number of functional elements	100	49

doi:10.1371/journal.pcbi.1000682.t001

Update of the parameters controlling the mouse TRA/TRD locus utilization during rearrangements by an experimental approach

In order to build the model, we firstly required information about the physical position of the V and J genes in the TRA/TRD locus. These data were provided by IMGT (ImMunoGeneTics database; http://imgt.cines.fr/) and summarized in Table 1. In addition, we needed to define parameters such as the opening location (the position where the opening mechanism begins), the opening speed for the access to V and J genes and the opening duration. Two more parameters were added, a first maturation step in order to eliminate the TRD genes and an opening offset as we supposed a certain rigidity of the DNA chain, thus two genes placed very close from each other cannot rearrange together.

Determination of the opening speeds for the V and J regions in the thymus. The data given in the first column of the Table 2 were obtained from rearrangements at the genomic level in BALB/c mice during thymic ontogeny and resume results presented in a previous work [9]. These data gave us the ontogeny days where V and J genes were first seen rearranged. In conjunction with physical gene positions, we calculated opening speeds that describe the progression of the accessibility to rearrangements over the Vα and Jα regions. Concerning the Jα region opening speed, in Fetal Fay 18th (F18), rearrangements of V19 with J61 to J48 corresponded to an opening of the J locus of 14773 bp in 24 h, thus the opening speed associated to these 24 h

Table 2. J locus accessibility: J genes seen rearranged to V19 during ontogeny.

Gestation day [§]	J opening	Opening distance [#]	Maximal opening Speed
F18 to F19	J61 to J48	14773 bp	615 bp/h
F19 to F20	J47 to J20	27618 bp	1150 bp/h
F20 to D0	J19 to J9	8998 bp	375–750 bp/h *
D0	J8 to J2	7940 bp	333 bp/h

[§]Thymus from Fetal Day 18 (F18) to Day of birth (D0); data are analyzed from Figure 5 in Pasqual et al. [9].

[#]Length of the DNA sequence corresponding to the J opening.

*For F20, the opening speed has been estimated between 375 bp/h to 750 pb/h depending on the offset of maxima 12 hours (9000 bp/24 h or 9000 bp/12 h).

doi:10.1371/journal.pcbi.1000682.t002

period is estimated to 615 bp/h. The same analysis was applied on F19, F20, and D0 (Day of birth). These data, fully presented in Table 2, showed that during ontogeny the opening speed of the J locus varies slightly, between 375 bp/h and 1150 bp/h, corresponding to an average opening speed of about 713 bp/h$\pm$3$\times$396 bp/h with a 99.9% confidence interval.

For determining the Vα region opening speed, there were used single member V families at each V region extremity of the TRA/TRD locus. For instance, V1 and V2, the most distal from the J region, as well as V19, V20, and V21, located at the nearest extremity to the J region, were analyzed. We found rearranged proximal V genes from F18, while rearrangements of distal V1 and V2 genes (1300 Kb distant from the proximal V genes) were only detected from D0. Hence, the entire V region takes about 3 days to get wide opened, allowing us to estimate the overall "opening V-speed" as broadly 18 Kb/h (1300 Kb/(3$\times$24 h)) with a 99.9% confidence interval of about 18$\pm$3$\times$5 Kb/h.

V-J rearrangements in peripheral T lymphocytes. In order to complete the study of thymus repertoire, we extended the analysis of V-J rearrangements in peripheral T lymphocytes from spleen and lymph nodes of adult mice. We reported the J use of 254 sequences extracted from peripheral T lymphocytes that express V14 on their surfaces [10]. The V14 family is composed of 6 members spread from 1145 Kb to 492 Kb in the V region. The Figure 1 A shows the J use by the whole V14 family. This distribution is in accordance with the two Hot Spots reported by Rytkonen *et al.* [3,23], which are indicated by two arrows over the histogram. Indeed, according to Rytkonen *et al.*, the J use distribution presents a preference for the J genes located over two regions named Hot Spots. The first Hot Spot (HSI) is located between J59 and J48 which corresponds to the region controlled by TEA; the second Hot Spot (HSII) is situated between J31 and J22. In addition, the Figure 1 presents the profiles of J used by the proximal V genes (V14-1, V14-2, and V14-3 on Fig. 1 B) and by the distal V genes (V14D-1, V14D-2, and V14D-3 on Fig. 1 C). These two histograms show that the proximal V to proximal J associations appear more frequent than the distal to distal associations and that the two Hot Spots are observed. Regarding the J region use, HSI is well observed for proximal V genes, and the HSII is well observed for distal V genes [10,24]. After presenting these biological data, the results generated by the model will be exposed.

Computational modelling approach

Values of the Parameters. Simulations were performed using two opening speeds chosen within intervals closed to the experimental 99.9% confidence intervals. For V speed, SV $\in$ [0.35 Kb/h, 34 Kb/h] and for J speed, SJ $\in$ [0.4 Kb/h, 1.55 Kb/h] with a mean opening speed of about 18 Kb/h for the V region and 1 Kb/h for the J region. The opening location of the simulation was fixed between the V and J genes in order to access directly to the TRAV and TRAJ genes after the first maturation, which was set to allow the elimination of the genes coding for TRD genes (region encompassing TRDV1 to TRDV5). Issues obtained from the modelling which best fit experimental data indicated that firstly, the duration of the first maturation step has a mean value of 5 hours, secondly, the number of successive rearrangements is 3 or 4, and thirdly, the opening duration before each rearrangement is 24 hours. The entire duration for the process of successive rearrangements is then 72 h or 96 h, which is in accordance with our data from ontogeny analyses (Table 2). When rearrangements were simulated by pairs, in order to account for the synchronization between the two alleles of individual cells, identical results were obtained.

Validation of the model by comparing simulated with thymic experimental data. The results of the model simulations and its comparison with the thymic experimental data are presented on Figure 2. Firstly, we present the global V and J uses in the simulated population (Fig. 2, A and B). The simulation program provides frequencies of every V to J genes associations in a matrix form. The columns display the J genes and the rows the V genes. Every intersection column/row indicates the frequency of the considered V to J association. It is then possible to sum all the V-J frequencies where a single V gene is implicated by adding the corresponding entire row. By doing this with all of the V genes, the global V region utilization is calculated (Fig. 2 A), and the sums of every matrix columns result in the global J gene utilization (Fig. 2 B). Overall, the uses of the V genes decrease from proximal to distal V genes, and the J region uses decrease from proximal to distal J genes. These tendencies were experimentally observed from thymic [9] and peripheral data as well [7,13].

Afterwards, we proceeded to quantify V1 and V21 rearrangements with a set of 9 J genes scattered along the J region in the thymus and compared these data with simulated outputs. Both experimental and modelling data show firstly that V1, which is the most distal V, has a low utilization rate of the proximal J genes and rearranges essentially the middle and distal J genes (Fig. 2, C, E, and G) and secondly that V21, which is the most proximal V, is mainly rearranged with the proximal J genes (Fig. 2, D, F, and H), following a Poisson distribution. The global utilization of J by V1 is similar in both experimental and simulated data. The frequencies from modelling are in correlation with the experimental nine J gene use (taken as representatives of the J region) by V1 and V21. In conclusion, the correspondence between data obtained by experimental analyses of rearrangements and those generated by *in silico* rearrangements validates the simulation program as model.

Program interface and generated graphics. The simulation program with its user interface (Fig. 3 A) provides a 2-dimensional diagram showing a conditional V-J rearrangement distribution for different V from proximal, central and distal positions (Fig. 3 B). Moreover, a 3-dimensional histogram of V-J rearrangements representing the all TRα chain combinational repertoire can be generated (Fig. 3 C). The program offers as well an interesting graphic representation designed to plot the J region use by complex V families. For the V14 family, for example, the program displays the complete J use by all the 6 V members (Fig. 4 A).

Stimulation model for peripheral T cell repertoire. In order to verify if the simulation model can be applied as a tool to determine the use of the different J by given V, we further quantified the use of 9 J by the 6 members of the V14 family in peripheral T lymphocytes. The resulting 54 V-J combinations, well spread along both the V and J regions, were plotted over the global V-J association 3-D graphic given by the simulated results (Fig. 5). For more precision, we plotted these frequencies of experimentally determined rearrangements over a fitted by 3-cubic spline fitting (Fig. 5 A) and a not fitted (Fig. 5 B) surfaces that interpolate the simulated frequencies. The experimental points over the surface appear in red, the ones under the surface appear in green. Beyond this visual adequacy, we demonstrated the accordance between simulated and experimental data by a numerical approach. We distinguished proximal, central and distal parts on the J region. For each of these parts, we compared the percentage of V14 rearrangements from experimental versus simulated data (Table 3). The V14 combinations with J61 to J48 represent 37% of the experimental data, and 35% of the simulated data, those with J47 to J24 stands for 46% and 50%, and

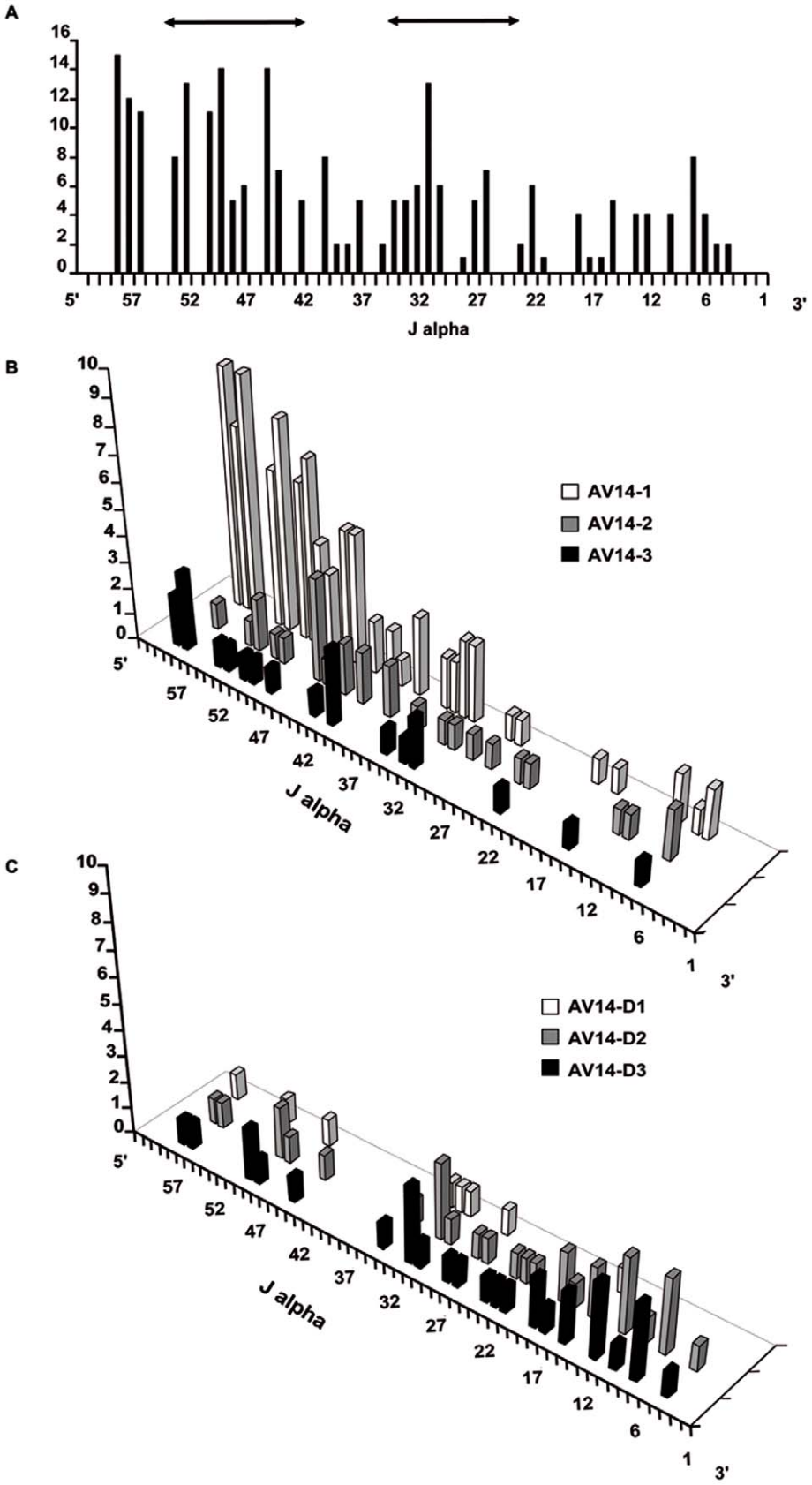


Figure 1. Quantification of the J region use by the V14 family. 254 V14 rearrangements were cloned from T lymphocytes, the V14 members and J genes were determined by sequencing [10]. (A) Profile of the J use by the six members of the V14 family. The two arrows indicate the localization of the two Hot Spots. (B) Profile of the three members the nearest from J genes and (C) J use by the most 5′ V14 members.
doi:10.1371/journal.pcbi.1000682.g001

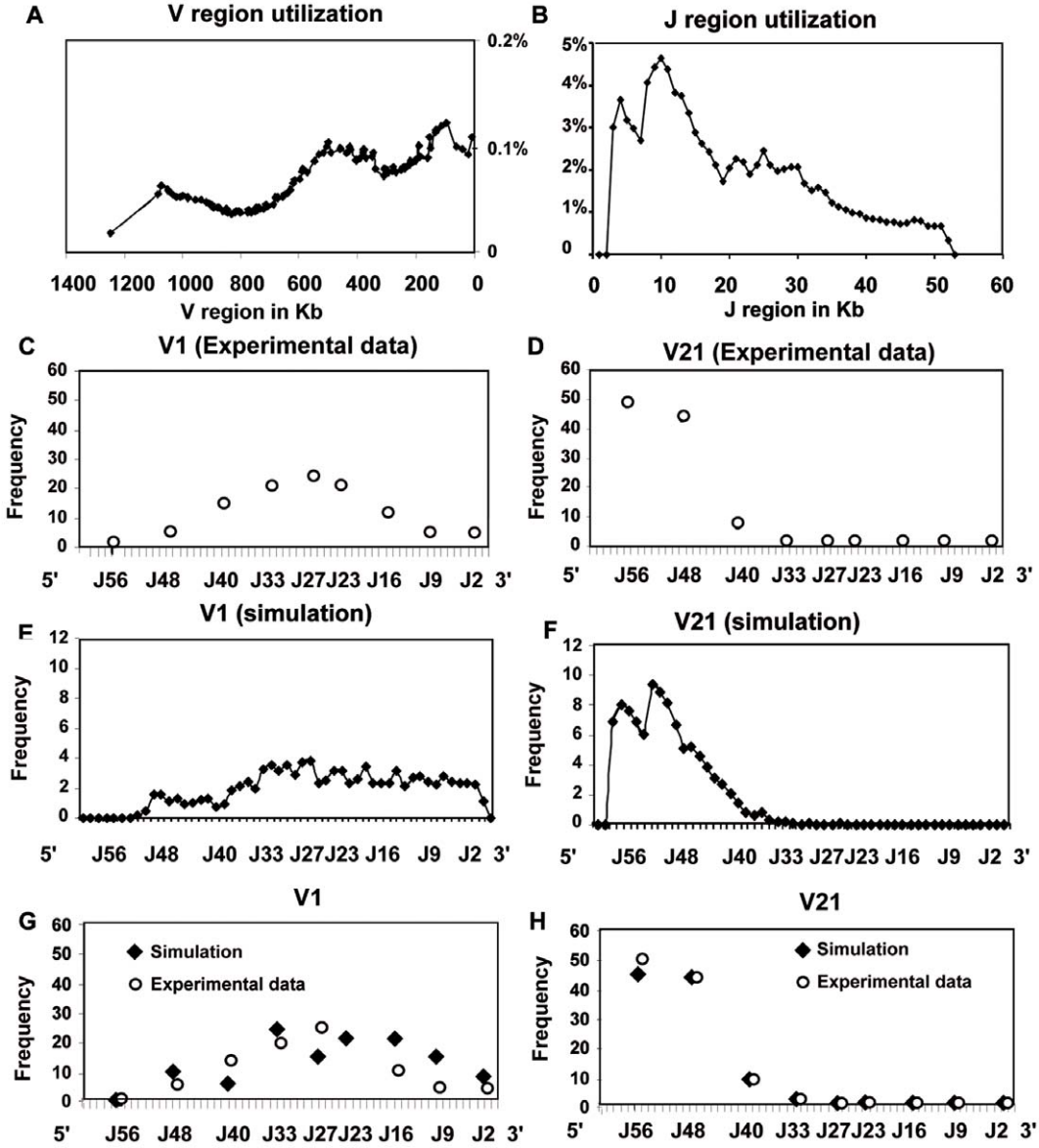


Figure 2. Validation of the modelling approach: analysis of the V and J region uses. (A) V region utilization: the X axis represents the V region in Kb. The Y axis shows the V gene percentage utilization in simulation. The simulated data sets have been normalized in order to be compared according to the following formula $X = (x - \text{average})/\text{Std deviation}$. The fixed parameters of the simulation were as follow, one million of alpha chains, ongoing 1 to 4 rearrangements with opening speeds of 18 Kb/h and 1.03 Kb/h for the V and the J region respectively; (B) J region utilization: the X axis represents the J region in Kb; (C) and (D) Amplitude of J region utilization by opposite V genes, V1 (distal) and V21 (proximal). The X axis represents experimental quantification on 9 J genes. The Y axis shows the experimental utilization frequency of 9 J genes by the V1 and V21 genes. (E) and (F) Amplitude of J region utilization in the model. The X axis represents the J genes. The Y axis shows the model frequency utilization by each J genes. (G) and (H) Superposition of experimental and simulated data for the 9 J genes. The X axis represents experimental quantification on 9 J genes. V and J regions utilization from simulated data are similar to experimental data obtained from [9].
doi:10.1371/journal.pcbi.1000682.g002

combinations with J23 to J1 represent 17% and 15% of experimental and simulated data respectively. Noticeably, when the sums of simulated rearrangements are extended to all V, the uses of the J localized in the proximal, central and distal parts of the locus are in the same range than those found with V14 genes (last column of the Table 3). Additionally, the J use by the whole V14 family (Fig. 4 A) from our simulated repertoire presents a distribution where two Hot Spots are clearly visible. These Hot Spots are in accordance with our own observation from peripheral T-lymphocytes (Fig. 1 A) and the experimental data of Rytkonen *et al* [3,23].

New insights in the VJ repertoire given by the model simulations. Based on our modelling study, four features concerning the V-J rearrangement process emerge. First of all, concerning the frequencies of the associations, the main information supplied by the model is that 96% (4704 out of 4900) of the V-J associations are probable. Hence, two areas where V-J combinations rarely occur could be defined (Fig. 3 C): the "A" area as associations between proximal V and distal J and the "B" area as associations between distal V and proximal J. These occasional associations are a consequence of a non-synchronized availability for rearrangements of the concerned V and J genes, as

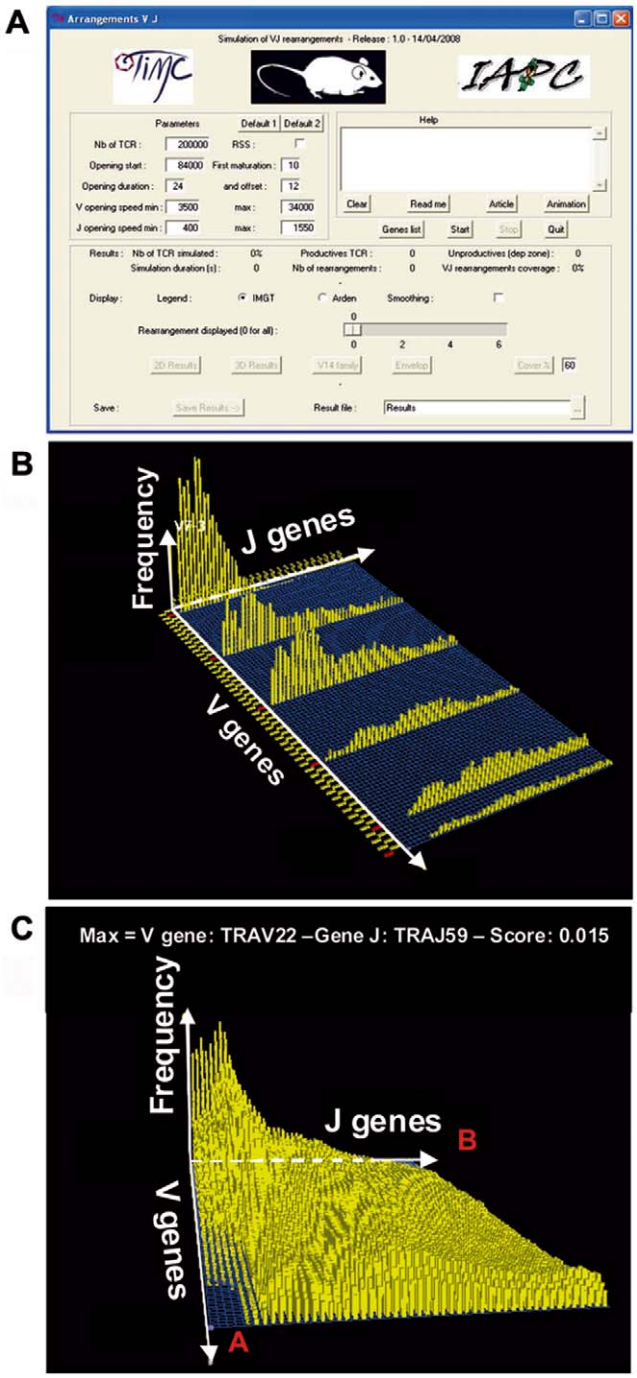


Figure 3. Model interface and results. (A) The main user interface window of the simulation program, (B) 2D representation of the rearrangement frequencies, (C) 3D representation of the rearrangement frequencies over all V and J gene associations.
doi:10.1371/journal.pcbi.1000682.g003

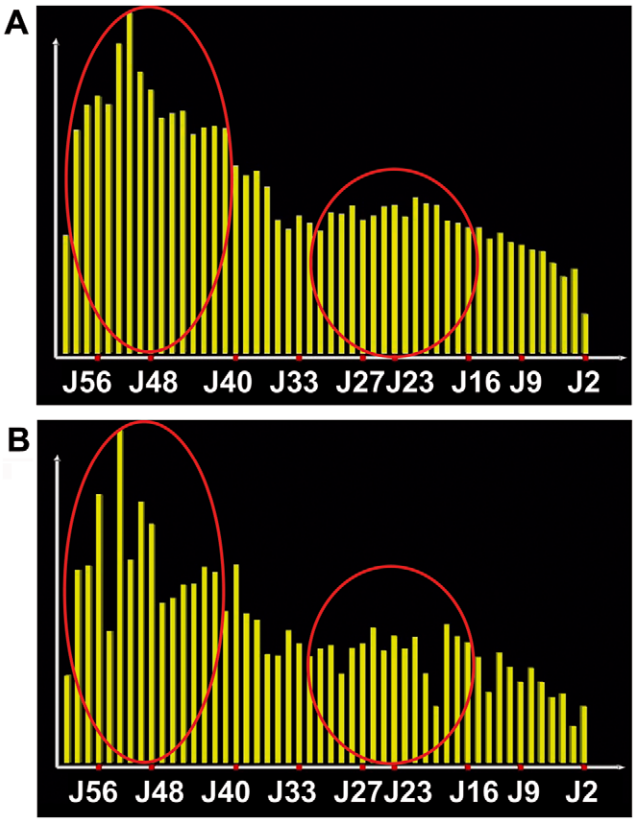


Figure 4. Representation of the V14 family rearrangement frequencies. Y axis represents the cumulated frequencies of all V14 genes with the J genes presented on the X axis, (A) without correction for RSS scores, (B) with correction accordingly to RSS scores. The two red ellipses show the localization of the two Hot Spots of recombination.
doi:10.1371/journal.pcbi.1000682.g004

already documented in experimental data [7,13,19]. Secondly, the model gives estimation for the frequencies of each V-J combination building the whole combinatorial repertoire shape. In the third place, the model states the influence of the Recombination Signal Sequence (RSS) on the V-J association distribution, which remains until now debated [25,26]. For that purpose, we ran simulations with and without taking into account the RSS scores (available in IMGT). The 2D graphs of the V14 family repertoire (Fig. 4, A and B) show that introducing variations accordingly to the RSS scores (Fig. 4 B) does not drastically affect the shape of the global repertoire distribution but leads to a local effect on certain J genes. The algorithm for choosing the J genes regarding their RSS score values has been favorably tested by using the Monte Carlo method. In the fourth place, the model provides information on the contribution of each wave of successive rearrangements to account for the total of V-J associations. Therefore, we tested the occurrence of 1 to 6 successive rearrangements in simulations. With 1 or 2 rearrangements, only the proximal V to proximal J associations are observed. With 5 or more rearrangements, the repertoire presents a distal border effect, corresponding to numerous rearrangements of the distal V and distal J regions, which are incoherent with experimental data. In conclusion, the overall repertoire generated by the simulation is in accordance with experimental data only by allowing 3 to 4 rearrangements, with a delay of 24 hours per rearrangement. Moreover, the contribution of each wave of successive rearrangements appears to decrease accordingly to their rank; 40% of the overall V-J associations is produced by the first wave of rearrangements, 33%, 19% and 8% come from the subsequent rearrangement waves respectively (Table 4).

Robustness of the model. To assure that the sampling size used in simulations was sufficient, we checked the representativeness of the repertoire by making sets of simulations ranging from 10^2 to 1.5×10^6 rearrangements. We found that

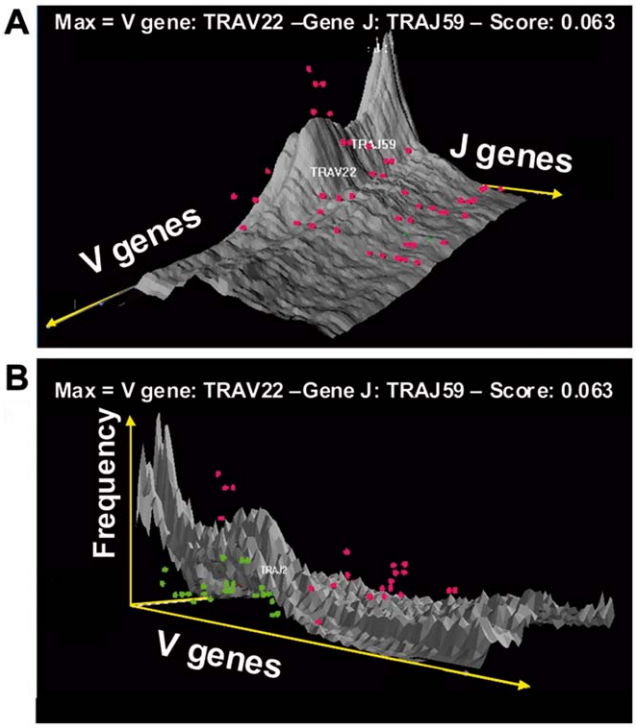


Figure 5. 3-D superposition of V14 family rearrangements. (A) The fitted simulated data and (B) non fitted simulated data are shown in grey shapes. The experimental points above the simulation shape are represented in red. The experimental points under the simulation shape are represented in green.
doi:10.1371/journal.pcbi.1000682.g005

diversity became relevant when the population size was higher than 5×10^5 rearrangements. This showed the pertinence of a repertoire calculation based on a 10^6 alpha chains population. A combinatorial diversity graph is plotted in Figure 6. Variations of about 5 to 10% in the values of the parameters (such as the intervals of the opening speeds Sv and Sj, opening duration, and offset) provided simulation results in concordance with the experimental data. However, larger variations in the values of the parameters induced major deviations (not shown) on the modelling simulation results compared to experimental data.

Finally, the consistency between simulated data and the frequencies observed in the thymus and the periphery validates our model as a relevant tool accounting for the mature repertoire of TRA/TRD.

Table 3. J region use by V14: comparison between experimental and simulation data.

	Experiment	Simulation	
	V14 *	V14 #	All V #
J61 to J48	37%	35%	33%
J47 to J24	46%	50%	51%
J23 to J1	17%	15%	16%

*Frequencies of rearrangements of V14 genes were calculated from Figure 2 in Aude-Garcia *et al.* [10], for the combinations with three J panels, corresponding to series of J genes scattered along the J region.
#Frequencies of rearrangements of V14 genes and of all V genes were calculated from modelling data for the combinations with same series of J genes.
doi:10.1371/journal.pcbi.1000682.t003

Table 4. Contribution of each rearrangement round into the total V-J combinatorial repertoire.

Rearrangement	First	Second	Third	Fourth
Percentage	40%	33%	19%	8%

doi:10.1371/journal.pcbi.1000682.t004

Discussion

This article focuses on a new approach to account for the features of the V to J rearrangement process in the TRA/TRD locus as well as to give a first estimation of the combinational repertoire in a 3-dimension representation. To accomplish this purpose, we have defined a mathematical model fitting experimental observations obtained from T lymphocytes rearrangements in the thymus. Jointly, the experimental data and the mathematical model made possible the interpretation of the mouse T-cell alpha chain repertoire characteristics. The evolution of the shapes for the V-J rearrangement frequencies in the simulations, presented in Figure 7, showed a transient bi-modal shape corresponding to the Hot Spots of V-J recombinations as observed in our experimental data and literature [3,10]. Furthermore, the model results also fit with V-J rearrangements obtained from T lymphocytes of the periphery. Therefore, our model provides a major improvement to previous attempts of simulation of the TRA combinatorial repertoire building [27].

Opening velocities and gene density. Our model is based on the fact that V and J regions are used in a progressive and decreasing manner from 3′ to 5′ for the V region and from 5′ to 3′ for the J region. Our quantitative approach points out that the physical position of the genes is the main structural parameter governing the uses of V and J genes. The J and V regions become accessible from proximal to distal genes according to an average "opening speed" of approximately 1 Kb/h for the J region and around 18 Kb/h for the V region. Interestingly, this difference between the V and J region opening speed values can be related to the gene density. In fact, the number of genes and the size of the V and J regions differ significantly. As a matter of fact, the V region length is 20 times larger than the J one (~1300 Kb versus

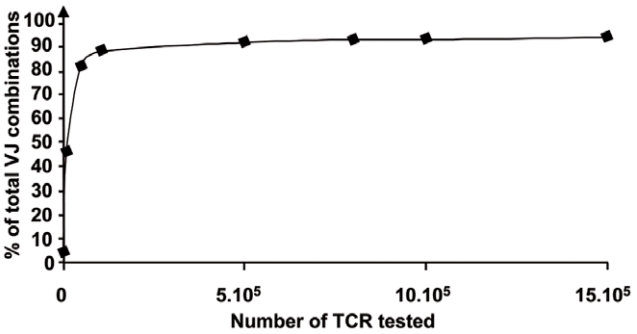


Figure 6. Combinational diversity of V-J combinations and population size. X axis represents the number of TR tested in the simulation, and Y axis indicates the percentage of the number of the different V-J combinations obtained by the simulation over the total number of possible V-J combinations.
doi:10.1371/journal.pcbi.1000682.g006

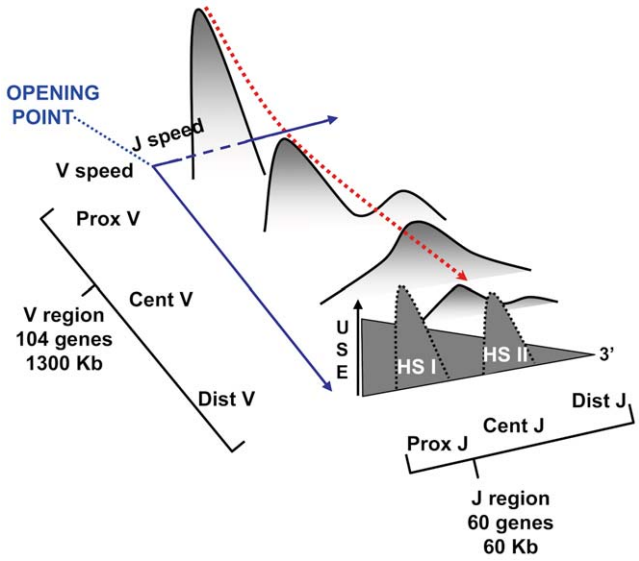


Figure 7. Schematic representation of the TRA/TRD locus use. The scheme shows the rearrangement distribution of 4 V genes with all J genes. The dashed red arrow indicates the decreasing frequency of rearrangements which correlates with the associations of distal V and J genes. A distal V gene is very rarely rearranged with proximal J genes because of the high recombination frequency between proximal V and J genes, leading to the proximal J genes deletion. 1) the first step governing the TRA/TRD locus utilization is defined by the opening of the most proximal V-J region, 2) According that T cells can undergo 1 to 3 secondary rearrangements, the second step giving the V and J accessible windows, is defined by the opening speeds of V and J regions. V speed is about 18 Kb/h whereas J speed is about 1 Kb/h, 3) J region has two Hot Spots of rearrangement. HSI is centred on J48 and HSII is centred on J30.
doi:10.1371/journal.pcbi.1000682.g007

~60 Kb), but the density of J genes is about 10 times higher than the density of the V genes (1 J by Kb versus 0.13 V by Kb). Consequently, the opening speeds calculated in terms of genes per hour is 1.4 for the V region and 0.83 for the J region (calculated as follows: the number of genes in the locus times the speed in Kb per hour, then divided by the length of the locus). Finally, taking into account the gene density of each region, the opening speeds of the V and J regions are almost identical. Our observations reinforce two putative scenarios being previously proposed to explain the opening of the J and V regions respectively. The first one, the J local service scenario was proposed for the J region [28]. It consists in a J use that follows small steps during the successive rearrangements. This local service is controlled by promoter activities associated to some of the other J genes. The second one, the V region express service proposes a large window for the gene accessibility to rearrangements controlled by enhancer activity. It reflects the larger utilization of a V region, whose genes are more scattered than those in the J one [9]. In addition, the speed calculations also take into consideration the regulatory elements controlling the gene accessibility [29]. As a remark, a parametric study with different speed values indicates that the proposed speeds are the only ones allowing a use of V and J regions correlating precisely with the experimental results.

Combinatorial repertoire distribution. Given that the totality of the V-J combination frequencies is computable in the framework of our model, we can visualize the whole simulated TRA combinatorial repertoire and thus estimate each V-J association frequency. Indeed, the probability of any combination is given by selecting a specific V-J combination on the 3D graph (Fig. 3 C).

Table 5. V-J association probabilities along the TRA locus.

	V21 Proximal	V4-2 Central	V2 Distal
J52 Proximal	0.148	0.050	0
J31 Central	0.002	0.029	0.023
J2 distal	0	0.003	0.007

9 V-J association probabilities given by the model. These results show an unbalanced use of the proximal and distal V and J genes. For instance, if all V-J combinations are equiprobable, the probability of each V-J association should be about 2.10^{-4}.
doi:10.1371/journal.pcbi.1000682.t005

Table 5 displays the probability of nine V-J combinations selected along the locus. Regarding the positions of the V and J genes over the locus, the probability of association varies between 0.01 for any V to distal J and 0.198 for any V to proximal J, highlighting the fact that a V to proximal J combination is about 20 times more probable than a V to distal J association.

Moreover, this 3D graphic points out that one central area in the V-J association plane contains the most represented combinations of the repertoire (Fig. 3 C). Furthermore, two areas (A and B in Fig. 3 C) reveal rarely represented V-J combinations, due to the fact that the V and J genes involved in these combinations are not accessible to get rearranged simultaneously. Concerning the A area, when the proximal J genes are recombining, the distal V genes are still inaccessible, and when they become accessible, the proximal J genes are deleted because of previous rearrangements. Similarly, on the B area, the non-synchronized accessibility of the proximal V genes and the distal J genes explains that their associations are not observed in the simulation.

The J region use confirms the existence of two Hot Spots. The progressive opening mechanism over the TRA locus provides V-J combinations that show a specific pattern; each given V gene is rearranged with a contiguous set of J genes. The distribution of these J genes presents a Poissonian distribution for the proximal V genes or a Gaussian shape for the distal V genes. The changes in the J use are progressive, depending on the V position: the more distal is a V gene, the more distal and larger is the set of J genes used, and the less represented are these V-J associations in the whole V-J repertoire (Fig. 3 C). There are two main probabilistic bases for the occurrence of the Hot Spots in the J region observed in the simulation results. The successive rearrangements are achieved by consecutive random choices of J genes, considering the progression of their access to recombination and optionally their RSS score values. The first J gene choice, corresponding to the first rearrangement, follows a Poissonian law whereas individually the two other J gene choices follow a Gaussian law (Fig. 8). Altogether, the consecutive random choices of J genes build a multimodal curve of occurrence, which allows the appearance of two Hot Spots (Fig. 8, solid line). The first Hot Spot results from the Poisson's distribution of the first J choice and the second one from the Gaussian distributions of the subsequent J choices. Moreover, it is important to remark that the density of J genes in the TRA locus is not uniform: J genes are less dense between J58-J47, J39-J28, and J14-J4, whereas they have a higher density between J45-J40 and J24-J15 (Fig. 9). The J gene density reaches its maximum in the area around J21-J22, corresponding to the place of the second Hot Spot. Interestingly, we observed the Hot Spots when the model parameter values were in the range where the opening dynamics allowed more frequent rearrangements within the areas of maximal gene density.

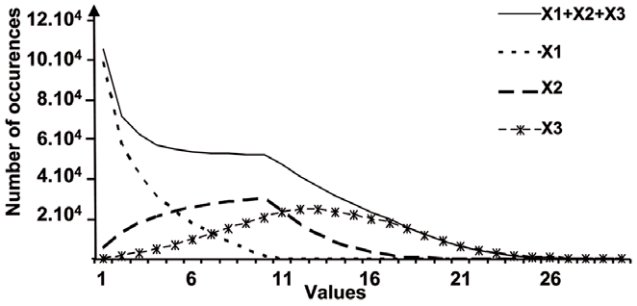


Figure 8. Successive rearrangements and building of the combinatorial repertoire shape. Three successive draws of random integers were done successively, the first one giving an integer x1 between 0 and 10 following a Poissonian law. The second and the third ones follow a Gaussian law, the second giving an integer x2 between x1 and x1+10, and the third giving an integer x3 between x2 and x2+10, and that 300 000 times. The first, second and third curves were added to build the sum curve.
doi:10.1371/journal.pcbi.1000682.g008

Number of secondary rearrangements. Secondary V-J rearrangements of the TRA/TRD locus are widely accepted [20,21,28,30,31]. However, the number of plausible secondary rearrangements remains unknown. Our model predicts that after the first maturation step of the TRA/TRD locus, consisting in the elimination of the TRD genes, the first V-J rearrangement of TRA is followed by a maximum of three secondary rearrangements. Nevertheless, each round of rearrangement contributes differently in the building of the whole repertoire, decreasing after each wave, and consequently, the fourth rearrangements have a weak contribution of 8%. This estimation of 4 total rearrangements is based on a realistic model including opening speeds from ontogenic experiments and may be more precise than Warmflash theoretical model's results that proposes a higher number of successive rearrangements [27]. It is important to consider that the number of secondary rearrangements can be affected by the lifespan of the rearranging T lymphocytes. For instance, the RORγ-deficient mice, presenting a shorter lifespan of DP thymocytes, show essentially proximal J genes rearranged, while Bcl-xL–transgenic mice, having DP with a longer lifespan, present a higher rate of distal J genes rearranged [28].

RSS influence. Regarding the Recombination Signal Sequence (RSS) influence, our model is able to incorporate the RSS diversity information through scores. The simulations using these RSS scores show a local quantitative influence but do not change the global profile of the frequency curves (Fig. 5, A and B). In conclusion, the RSSs may only influence local specificities within the accessibility windows moving across the TRA/TRD locus in a bi-directional way. This is in good accordance with mouse TRB locus observations showing that V gene RSSs neither correlate with any particular restriction of J genes nor with any high V-J rearrangement frequencies [32].

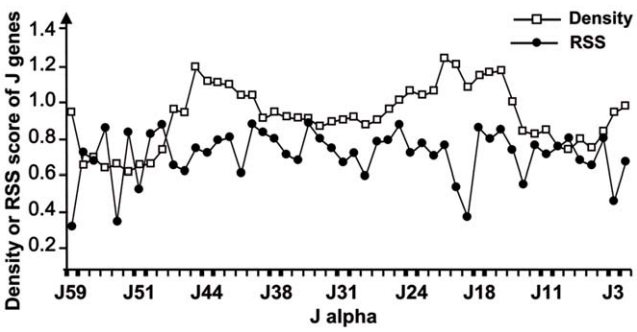


Figure 9. Density and RSS scores of the J genes. Values for the density (open squares) and RSS scores (dark circles) were calculated, as described in methods, for each J gene from the four previous and next genes. X axis represents the J genes, the Y axis the density or the RSS score for all J genes.
doi:10.1371/journal.pcbi.1000682.g009

The sequential windowing model: a tool to determine the peripheral V-J association frequencies. We previously observed experimentally and tested statistically that the thymic and the peripheral repertoires showed similar profiles of J uses by the V14 family [10]. When we compared the experimental data of the uses of J genes by the V14 family members we found that the J use profiles fitted our model results. It is acknowledged that the V14 is a multimember family and may be representative of the J use by different V genes. Finally, our data suggest that the model would be used as a tool to determine the V-J association frequencies in the peripheral T lymphocytes.

All these remarks support the realistic character of our model, which includes the essential features of the V-J rearrangement process in the TRA/TRD locus. In conclusion, the combination of experimental and mathematical approaches gives new insights on combinatorial repertoire biases due to non-equiprobable V-J combinations in TRA/TRD rearrangements, and allows defining more accurately the TRA/TRD primary combinatorial repertoire. In the future, the model could be adapted to other loci and other species, to propose accurate estimations of the V-J combinatorial diversity, giving a dynamical vision of the immune diversity generation during differentiation of T cells and B lymphocytes.

Materials and Methods

Nomenclature. Official nomenclature for V and J genes is chosen according to the IMGT database (http://imgt.cines.fr). NCBI (National Center for Biotechnology Information) accession numbers are AE008683-AE008686 for the mouse V region and M64239 for the J region. Positions of each V and J genes were calculated based on these data as previously described [6].

Mouse. BALB/c mice were purchased from Charles-River (L'Abresles, France). Mice were housed and humanely killed according to relevant national guidelines. No experimental work was done on living animals. Fetal thymi were obtained from timed pregnancies, where Fetal Day 1 (F1) corresponds to the day of detection of a vaginal plug. Thymic lobes from embryonic mice were pooled and mechanically dissociated in PBS before DNA extraction.

Multiplex PCR assay analysis. multiplex PCR assays and quantification analysis were done as described in [9,33,34].

Quantitative PCR. Real time PCR were performed on a Light CyclerTM (Roche diagnostics, Meylan, France). The specificity of the unique amplification product was determined by melting curve analysis and using migrations on agarose gels followed by southern blotting with the corresponding internal V probes [9].

Description of the computational model

Our **sequential windowing model** is a specific model of successive windows, each step corresponding to a differential motion of the window extremities, with opening velocity faster on the V side than on the J one. The simulation of the V-J rearrangements in the TRA/TRD locus is based on a computational occurrence discrete model using parameters determined

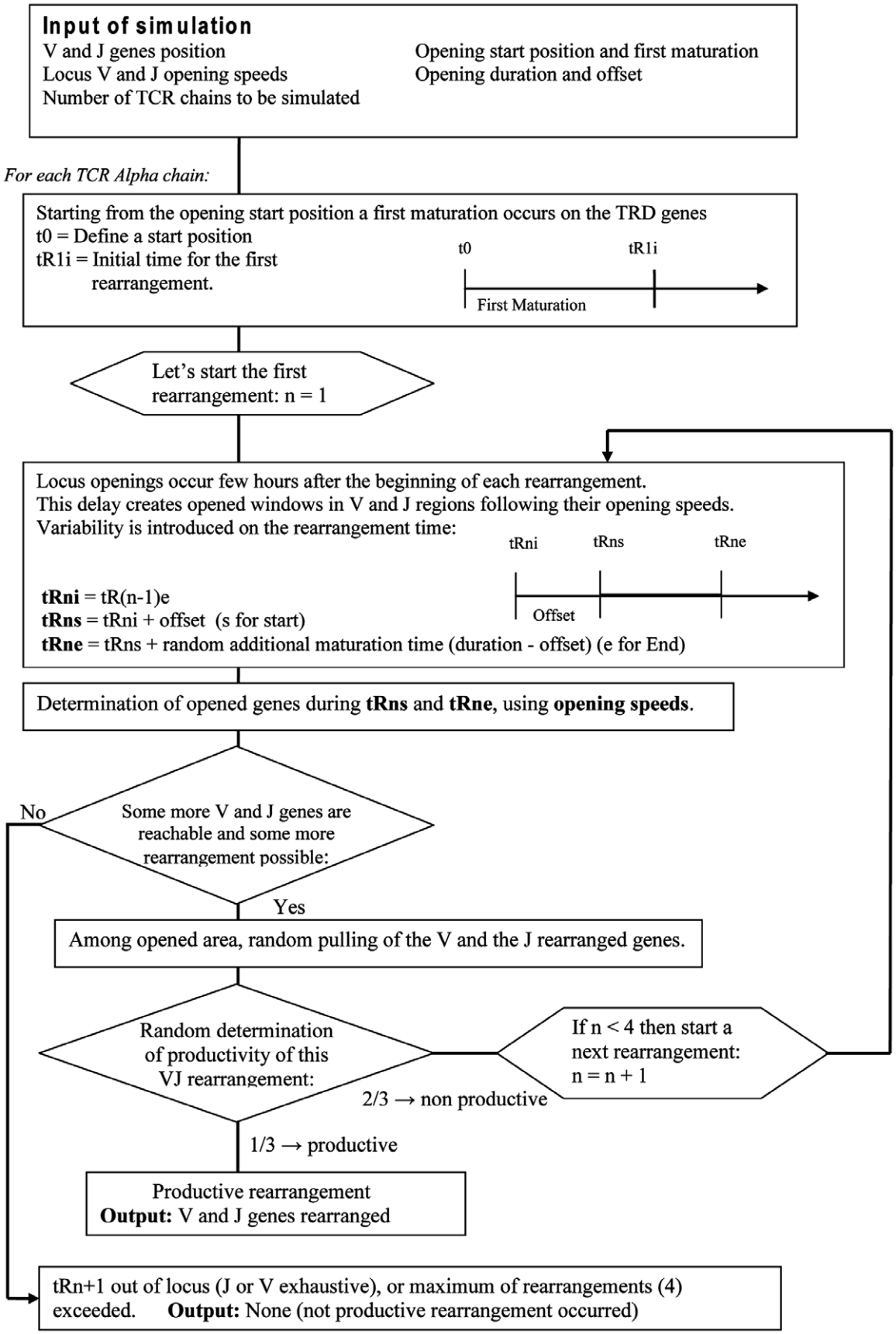


Figure 10. Flow diagram for the sequential windowing model algorithm.
doi:10.1371/journal.pcbi.1000682.g010

from experimental data (Fig. 10). The model consists in dynamical rules depending on constant (structural) and experimental parameters. The constant parameters are the physical positions of the 100 V and 49 J functional genes and the first TRA/TRD locus opening points. The varying parameters (whose sensitivity will be studied by simulation) are the opening speed intervals of the V and J regions, the opening duration before each rearrangement, and the opening offset.

We define variables and dynamical rules of the model as follows:

1) **The physical positions of V and J genes in the TRA/TRD locus** are calculated by measuring the genomic distance from the TRAC chain.

2) **The first maturation step** is fixed before the first rearrangement allowing the deletion of the region encompassing TRDV1 to TRDV5 including the TRD locus. Its duration value is a constant parameter, which was determined by simulation varying values between 0 hours and 10 hours.

3) **The opening location** describes the site where the opening mechanism begins. This site is fixed at the TEA location [35].

4) **The opening speeds** of the V and J genes calculated above are denoted respectively by S_V and S_J. They are random variables between a minimum and a maximum. Each TCRA locus is simulated independently to be consistent with the absence of allelic exclusion. Furthermore, rearrangements were also simulated by pairs using the same V and J opening speeds, in order to account for the synchronization between the two alleles of individual cells.

5) **The Opening duration before each rearrangement** is a constant parameter whose value was determined through simulations, by varying values between 2 hours and 50 hours.

6) N denotes the number of authorized **secondary rearrangements** during the simulation. It is a varying parameter as well, values between $N=0$ to $N=6$ were studied by simulations.

7) The **probability to perform an in-frame rearrangement** at any step k ($1 \le k \le N$) is fixed to 1/3 that is the maximal possible value. If the rearrangement is randomly determined in-frame, the procedure is stopped for this locus and the V-J association generated is stored. If the simulation gives an out-of frame rearrangement at the step k, a new secondary rearrangement is randomly executed at the step k+1 on the available part of the locus. The new window of accessibility is calculated in base of the "**opening speeds**" and the **"opening duration before each rearrangement"** parameters. This successive rearrangement procedure remains until either an in-frame rearrangement occurs or k equals the maximum number of rearrangements N.

8) We refer to the length of the window of accessible DNA over the V region at a step k as LV_k (and LJ_k for the J window length), These windows progress from the proximal to distal extremities of the TRA/TRD locus V and J regions. The LVs and LJs verify the equations:

$$LV_1 = S_V(t_0+\tau_1), \ldots, LV_k = LV_{k-1} + S_V(\tau_k), fork \ge 2,$$

$$LJ_1 = S_J(t_0+\tau_1), \ldots, LJ_k = LJ_{k-1} + S_J(\tau_k), fork \ge 2$$

where the opening offset time t_0 denotes a minimal time of opening and τ_k ($k \ge 1$) are random variables uniformly distributed between t_0 and the end of the opening process.

9) For every rearrangement occurrence, we define for each gene V_i (J_j respectively) a Boolean variable BV_{ik} (resp. BJ_{jk}) equal at the k^{th} rearrangement to 1 if the gene is open ("accessible"), and to 0 if it is closed ("non-accessible") or if it has been deleted during a previous rearrangement of order i ($i<k$).

10) The RSS score KV_i (resp. KJ_j) represents for each gene V_i (J_j respectively) the homology percentage compared to a consensus sequence. This score basically takes value p, $0 \le p \le 1$, if there is p % of identity between the RSS and the consensus proposed by Glusman *et al.* [25], allowing us to estimate a RSS identity score ranging from 0.3 to 1, where 1 corresponds to a fully consensus RSS (see additional data). Our RSS score is in agreement with the status of functional versus pseudo V or J gene ($0.3<$ pseudo rearrangement score <0.65; $0.65<$ functional score <1). According to its non-functional status, a J-pseudo recombination is never found rearranged and consequently the corresponding J-RSS score is assimilated to zero in simulations. RSS score is equal to 1 when simulation is done without taking account the RSS. We call FV_k (resp. FJ_k) the distribution function (i.e., the relative length) obtained after the k^{th} rearrangement by adding the BV (resp. BJ) variables:

$$FV_k(i) = \left(\sum_{m=1,\ldots,i} KV_m BV_{mk}\right) \Big/ \left(\sum_{m=1,\ldots,104} KV_m BV_{mk}\right)$$

$$FJ_k(j) = \left(\sum_{m=1,\ldots,j} KJ_m BJ_{mk}\right) \Big/ \left(\sum_{m=1,\ldots,60} KJ_m BJ_{mk}\right)$$

11) At step k, we choose the distribution functions FV_k and FJ_k corresponding to the random variables RV_k (resp. RJ_k) uniform on [0,1] and we calculate a number NVk (resp.NJk) equal to $\inf(FV_k^{-1}(RV_k))$ (resp. $\inf(FJ_k^{-1}(RJ_k))$). NVk and NJk corresponding to the V and J genes to rearrange.

The **number of simulated TRAD loci** gives the size of the simulated population. In the figures shown in this paper, 1 million of V-J rearrangements have been simulated.

The **simulation Output** is presented in a matrix form incremented by the successive V-J in-frame rearrangements. Final results show the total number of V-J combinations available at the end of the whole simulation. These results can be plotted in different 3D representations using the interface. It is also possible to display the results for multi member V famillies corresponding to the real time PCR's.

Author Contributions

Conceived and designed the experiments: JD PNM EJM. Performed the experiments: FT MAS NP AD TPB. Analyzed the data: FT MAS OH NP TPB VH JD EJM. Contributed reagents/materials/analysis tools: OH AD TPB VH. Wrote the paper: FT MAS OH NP JD PNM EJM.

References

1. Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM (2006) Accessibility control of V(D)J recombination. Adv Immunol 91: 45–109.
2. Jouvin-Marche E, Aude-Garcia C, Candeias S, Borel E, Hachemi-Rachedi S, et al. (1998) Differential chronology of TCRADV2 gene use by alpha and delta chains of the mouse TCR. Eur J Immunol 28: 818–827.
3. Rytkonen M, Hurwitz JL, Tolonen K, Pelkonen J (1994) Evidence for recombinatorial hot spots at the T cell receptor J alpha locus. Eur J Immunol 24: 107–115.
4. Abarrategui I, Krangel MS (2007) Noncoding transcription controls downstream promoters to regulate T-cell receptor alpha recombination. Embo J 26: 4380–4390.
5. Villey I, Caillol D, Selz F, Ferrier P, de Villartay JP (1996) Defect in rearrangement of the most 5′ TCR-J alpha following targeted deletion of T early alpha (TEA): implications for TCR alpha locus accessibility. Immunity 5: 331–342.
6. Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, et al. (2004) IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Nucleic Acids Res 32: D51–54.
7. Thompson SD, Pelkonen J, Rytkonen M, Samaridis J, Hurwitz JL (1990) Nonrandom rearrangement of T cell receptor J alpha genes in bone marrow T cell differentiation cultures. J Immunol 144: 2829–2834.
8. Thompson SD, Pelkonen J, Hurwitz JL (1990) First T cell receptor alpha gene rearrangements during T cell ontogeny skew to the 5′ region of the J alpha locus. J Immunol 145: 2347–2352.
9. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, et al. (2002) Quantitative and qualitative changes in V-J alpha rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor alpha chain repertoire. J Exp Med 196: 1163–1173.
10. Aude-Garcia C, Gallagher M, Marche PN, Jouvin-Marche E (2001) Preferential ADV-AJ association during recombination in the mouse T-cell receptor alpha/delta locus. Immunogenetics 52: 224–230.
11. Krangel MS (2003) Gene segment selection in V(D)J recombination: accessibility and beyond. Nat Immunol 4: 624–630.
12. Oltz EM (2001) Regulation of antigen receptor gene assembly in lymphocytes. Immunol Res 23: 121–133.
13. Davodeau F, Difilippantonio M, Roldan E, Malissen M, Casanova JL, et al. (2001) The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. Embo J 20: 4717–4729.
14. Sleckman BP, Bardon CG, Ferrini R, Davidson L, Alt FW (1997) Function of the TCR alpha enhancer in alphabeta and gammadelta T cells. Immunity 7: 505–515.
15. Hawwari A, Bock C, Krangel MS (2005) Regulation of T cell receptor alpha gene assembly by a complex hierarchy of germline Jalpha promoters. Nat Immunol 6: 481–489.
16. Hawwari A, Krangel MS (2007) Role for rearranged variable gene segments in directing secondary T cell receptor {alpha} recombination. Proc Natl Acad Sci U S A 104: 903–907.
17. Mauvieux L, Villey I, de Villartay JP (2003) TEA regulates local TCR-Jalpha accessibility through histone acetylation. Eur J Immunol 33: 2216–2222.
18. Marche PN, Kindt TJ (1986) Two distinct T-cell receptor alpha-chain transcripts in a rabbit T-cell line: implications for allelic exclusion in T cells. Proc Natl Acad Sci U S A 83: 2190–2194.
19. Krangel MS (2009) Mechanics of T cell receptor gene rearrangement. Curr Opin Immunol 21: 133–139.
20. Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, et al. (1993) Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. J Exp Med 178: 615–622.
21. Wang F, Huang CY, Kanagawa O (1998) Rapid deletion of rearranged T cell antigen receptor (TCR) Valpha-Jalpha segment by secondary rearrangement in the thymus: role of continuous rearrangement of TCR alpha chain gene and positive selection in the T cell repertoire formation. Proc Natl Acad Sci U S A 95: 11834–11839.
22. Lefranc MP (2001) IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 29: 207–209.
23. Rytkonen MA, Hurwitz JL, Thompson SD, Pelkonen J (1996) Restricted onset of T cell receptor alpha gene rearrangement in fetal and neonatal thymocytes. Eur J Immunol 26: 1892–1896.
24. Gahery-Segard H, Jouvin-Marche E, Six A, Gris-Liebe C, Malissen M, et al. (1996) Germline genomic structure of the B10.A mouse Tcra-V2 gene subfamily. Immunogenetics 44: 298–305.
25. Glusman G, Rowen L, Lee I, Boysen C, Roach JC, et al. (2001) Comparative genomics of the human and mouse T cell receptor loci. Immunity 15: 337–349.
26. Livak F, Burtrum DB, Rowen L, Schatz DG, Petrie HT (2000) Genetic modulation of T cell receptor gene segment usage during somatic recombination. J Exp Med 192: 1191–1196.
27. Warmflash A, Dinner AR (2006) A model for TCR gene segment use. J Immunol 177: 3857–3864.
28. Guo J, Hawwari A, Li H, Sun Z, Mahanta SK, et al. (2002) Regulation of the TCRalpha repertoire by the survival window of CD4(+)CD8(+) thymocytes. Nat Immunol 3: 469–476.
29. Osipovich O, Milley R, Meade A, Tachibana M, Shinkai Y, et al. (2004) Targeted inhibition of V(D)J recombination by a histone methyltransferase. Nat Immunol 5: 309–316.
30. Huang J, Muegge K (2001) Control of chromatin accessibility for V(D)J recombination by interleukin-7. J Leukoc Biol 69: 907–911.
31. Rytkonen-Nissinen M, Hurwitz JL, Pelkonen S, Levelt C, Pelkonen J (1999) Early activation of TCR alpha gene rearrangement in fetal thymocytes. Eur J Immunol 29: 2288–2296.
32. Wilson A, MacDonald HR, Radtke F (2001) Notch 1-deficient common lymphoid precursors adopt a B cell fate in the thymus. J Exp Med 194: 1003–1012.
33. Fuschiotti P, Pasqual N, Hierle V, Borel E, London J, et al. (2007) Analysis of the TCR alpha-chain rearrangement profile in human T lymphocytes. Mol Immunol 44: 3380–3388.
34. Mancini SJ, Candeias SM, Di Santo JP, Ferrier P, Marche PN, et al. (2001) TCRA gene rearrangement in immature thymocytes in absence of CD3, pre-TCR, and TCR signaling. J Immunol 167: 4485–4493.
35. de Chasseval R, de Villartay JP (1993) Functional characterization of the promoter for the human germ-line T cell receptor J alpha (TEA) transcript. Eur J Immunol 23: 1294–1298.

Annex 4

Numerical Model for the Vα-Jα Gene Use in Human TRA/TRD locus: Recombination Dynamical Rules. Thuderoz F, Hansen O, Simonet MA, Daris A, Borel E, Marche PN , Jouvin-Marche E, Demongeot J. (submitted).

Numerical Model for the Vα-Jα Gene Use in Human TRA/TRD locus: Recombination Dynamical Rules

Florence Thuderoz[1,2], Olivier Hansen[1,2], Maria-Ana Simonet[1,2,3], Aurélie Daris[2,3], Eve Borel[2,3], Patrice Noël Marche[2,3], Evelyne Jouvin-Marche[2,3,§], Jacques Demongeot[1,2,§].

1 CNRS, Laboratoire TIMC-IMAG, UMR 5525, Grenoble, France, 2 Université Joseph Fourier-Grenoble I, Faculté de Médecine, Grenoble, France, 3 INSERM U823, Institut Albert Bonniot, Grenoble, France.

Abstract

V(D)J recombination constitutes a somatic site specific DNA recombination originating lymphocyte antigen receptor diversity in jawed vertebrates. Concerning T-Cell receptor alpha-chains, V and J genes are used from the inside out of the TRA locus during successive rearrangements, with no allelic exclusion. Beside extensive quantifications of V-J associations from human thymic genomic DNA, a model approach is proposed. Comparison between model and experimental results enhances knowledge about kinetics and dynamical rules controlling human V-J segment use. Predictions are made about parameters not accessible through experiments, like the number of successive rearrangements (4 or 5) or the successive accessibility window sizes and locations. Finally, the precise progression of gene accessibility to rearrangements, according to non-constant opening speeds, together with a synchronized opening of the J regions between both alleles, are sufficient to fully explain both the experimental V-J frequencies currently available and the interallelic J usage. Model and experimental results provide a coherent representation of Vα-Jα combinatorial repertoire. Comparison with previous studies led on mouse draws a fine conservation of Vα-Jα rearrangement dynamics between both species.

Running title: Model for the Human TRA/TRD locus Vα-Jα Gene Use.
Keywords: human V(D)J recombination, TRA/TRD locus, Vα and Jα gene use, TRα repertoire Diversity.
Word count: 32000 characters (with spaces).
Subject categories suggested: Immunology, Bioinformatics.
§: These authors contributed equally to this work.
Funding: This work was supported by the institutional Grants from Institut National de la Santé et de la Recherche Médicale (INSERM), from Centre National de la Recherche Scientifique (CNRS), and by the EC Alfa project IPECA and NoE VPH. FT was supported by a fellowship from the Agence Nationale de la Recherche et de la Technologie, France. M-AS was supported by a fellowship from Région-Rhone Alpes "Cluster 10 and 11".

Introduction

Infectious agent multiplicity and dysfunctional host cells, like tumor cells, pushed jawed vertebrates to develop mechanisms permitting the production of an extensive variety of antigenic receptors (Rast *et al*, 1997). In humans and rodents, T cells, responsible of cell mediated immune response, mostly express clonotypic αβ T Receptors (TR) on their surface. The α and β protein chains are coded by loci being non-functional in the germ-line configuration: TRA locus includes multiple different copies of Variable (V) and Junction (J) genes, and a unique constant

(C) gene; TRB locus encompasses Diversity (D) genes as well. Several combinations of V(D)J genes are generated in developing lymphocytes by means of somatic site-specific DNA rearrangements (Bassing *et al*, 2002). The RAG, recombinase complex involved in the initiation of the V(D)J recombination, targets Recombination Signal Sequences or RSSs which are short sequences of 12 or 23 bp located next to each V, D and J genes (Oettinger *et al*,1990). The RAG catalyses double strand breaks between the RSSs and the genes; the DNA sequence previously present between the two genes to rearrange is deleted (Takeshita *et al*, 1989). Before the two genes are linked by ubiquitously expressed DNA repair proteins, some imprecision is introduced in the joining: random removal and non-templated addition of nucleotides increase the repertoire through junction diversity (Azuma *et al*, 1984; Gilfillan *et al*, 1993; Bleakley *et al*, 2008). Due to the non frank coding joints, a maximum of 1/3 of the performed rearrangements produce in-frame rearrangements, which conserve the proper translation reading frame of the gene (Coleclough *et al*, 1983). A last factor of TR αβ repertoire diversity consists in heterodimer pairing. If this step is successful, after surface expression of the TR, the T cells will undergo thymic selection, which guarantees only functional and non-autoreactive T cells will migrate to the periphery. At last, V(D)J rearrangements constitute a somatic neo-gene creation giving the organism the possibility to build a vast immune repertoire diversity.

Although profiles of TRβ repertoire are well established, knowledge related to TRα chain diversity remains restricted because few anti-VAD antibodies are available. Vα to Jα rearrangements take place during the CD4+CD8+ Double-Positive (DP) stage of T cell intrathymic development. The TRA locus is known to go through multiple rearrangement rounds (Petrie *et al*, 1993), without strict allelic exclusion (Krangel, 2009). When a Vα-Jα in-frame rearrangement is achieved in a cell, pairing with the β chain is assayed, leading to surface expression and positive selection that are both needed to allow the TR-mediated signals stopping further rearrangements. Vα and Jα genes are used sequentially from inside the locus toward distal genes. First rearrangements use Jα genes proximal to the Vα region (5') and Vα genes proximal to the Jα region (3'); successive secondary rearrangements implicate progressively more distal Jα and more distal Vα genes (Thompson *et al*, 1990; Jouvin-Marche *et al*, 1998; Aude-Garcia *et al*, 2001; Pasqual *et al*, 2002; Krangel *et al*, 2004). This inside out use depends on Cis regulating elements conserved between mouse and human: enhancers and promoters constrain accessibility over Jα and Vα regions. The TRα enhancer (Eα), located at the 3' end of the C gene, firstly activates two promoters that control the use of proximal J genes: T early α promoter (TEA) and J49 (Villey *et al*, 1996; Hawwari *et al*, 2005). The structure of the Jα region (TRAJ) is remarkably conserved between human and mouse genomes: 61 J genes (49 functional) are spread over a 71Kb J region and 60 J genes (44 functional) are spread over 64Kb respectively (see www.imgt.org for locus representation). The human Vα region includes 54 V genes (44-46 functional) spread over 800 kb, whereas in mouse, several duplications during evolution formed a longer V region (1400kb) encompassing 70 to more than 100 V genes depending on the haplotype (Gahéry-Ségard *et al*, 1996). The random V-J association model, taking into account non-functional genes, gave 2156-2254 V-J association possibilities for the human species. This value was reduced to roughly 2000 associations because of the inside out use of the TRA locus that leads to improbable associations of proximal V to distal J genes and of distal V to proximal J genes (Fuschiotti *et al*, 2007; Jouvin-Marche *et al*, 2009). Within these 2000 associations, knowledge about each individual V-J

association frequency remains to debate; V-J association frequencies constitute a further limitation of the combinatorial diversity that reduces the available TRα repertoire.

After recent advances in the field of TRA recombination modeling (Warmflash *et al*, 2006; Thuderoz *et al*, 2010), the present study addresses precisely Jα and Vα gene use in the human TRA locus on the basis of both experimental and model approaches. A number of 80 Vα-Jα associations were quantified in the human thymus using genomic real time PCRs. The corresponding 8 Vα and 10 Jα genes examined were chosen regularly placed along the Vα and Jα regions in order to offer a first experimental sampling of the human TRα computational repertoire in the thymus. Our successive windowing model (Figure 1), accounting for the Vα to Jα rearrangement process in mouse (Thuderoz *et al*, 2010) was adapted to the human locus structure particularities. Human model simulation results were compared with our experimental V-J quantifications and with interallelic J usage (Davodeau *et al*, 2001); the quality of the fit tells a specific and non constant progression of gene accessibility to rearrangements over the V and J regions along with a synchronized opening of the J region between both alleles is sufficient to fully explain V-J combinatorial repertoire features. Model and experiment results provide a complete representation of Vα-Jα combinatorial repertoire, which displays unbalanced V-J frequencies. The study showed that the dynamic rules and kinetics governing Vα to Jα rearrangements are practically conserved between human and mouse species. Interestingly, the model approach allows making some predictions about parameters, which cannot be determined through experimentation, like number of successive rearrangements or size and position of successive DNA accessible windows. Among other results, model studies predict the human locus undergoes 4-5 successive rearrangement rounds versus 3-4 rearrangements rounds in mouse. A supplementary rearrangement would give the human species one additional chance per allele to generate an in-frame rearrangement, which constitutes a major selective advantage compared to the mouse, compensating for the lower number of human V genes compared to that in mouse.

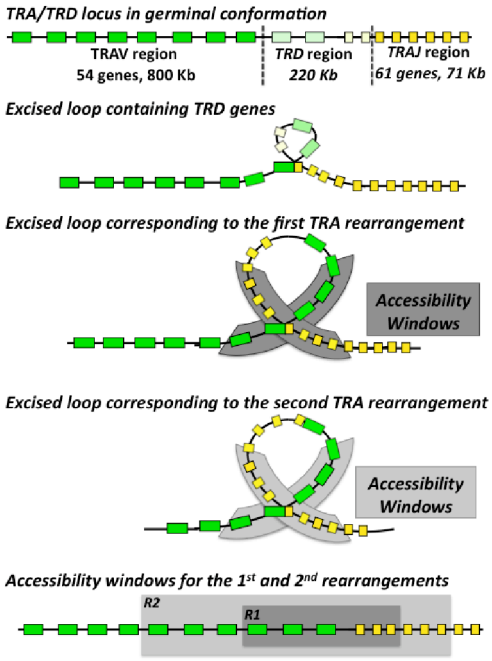


Figure 1: Successive windowing mechanism. After a first maturation allowing the excision of TRD genes, accessibility windows are calculated for each rearrangement, progressing over the V and J regions according to an inside out manner. Only the schemes corresponding to the first and second rounds of TRA rearrangements are represented and the last diagram provides a distinctive representation of the window progressions.

Results

Experimental Results

V-J association relative abundances in human thymi

Figure 2 presents relative measures of 80 V-J associations determined by quantitative genomic PCR analysis from 4 human thymic samples. Corresponding 8 V and 10 J genes appear ordered according to their physical

position over the locus. Amplification of G3PDH gene was used as a normalizer and results were expressed in arbitrary units, indicating the differences in cycle numbers at which PCR products were first detected. Average values and standard deviations are indicated for each V-J association and a color code illustrates V-J average levels. Results are consistent among the individual samples, with a maximum standard deviation value of 3.73 and an average standard deviation of 1.06. In detail, V41 and V40 genes, located next to the J region (3'end of the translated strand), rearranged predominantly with J genes from J61 to J24. The V38, V35, V30, and V 21 genes, located in the central part of the V region, mainly rearranged with J genes from J61 to J10. V16 used from J56 to J10 segments and eventually, V2 gene, situated in the distal part of the V region (3'end), rearranged mostly with distal J genes (3' end of the J region), from J33 to J5. All in all, Figure 2 clearly demonstrates a higher abundance of associations implicating V and J genes symmetrically placed in the germinal configuration of the TRA locus, namely proximal V to proximal J, centered V to centered J, and distal V to distal J genes.

Model Approach

Model interface and generated graphs

The simulation program based on the successive windowing model presents a convivial interface (Fig. 3) providing the user a handful way to vary the parameter values in order to test different scenarios for V-J recombination mechanisms. Simulation results are displayed on a 3-dimensional histogram of V-J rearrangements representing the whole TRα chain combinatorial repertoire. The programs offer as well a graphic representation designed to plot the J region use by some V genes by clicking on the V gene axis directly on an interactive graph. For distinctive usages, an Excel™ file is generated giving direct access to the simulated results: frequencies $F_{Vi\text{-}Jj}$ of all the V_i-J_j associations, details about the associations generated through each rearrangement round, couple of J genes used

<----------TRAJ REGION----------> (Proximal J Genes (5' end) … Distal J Genes (3' end))

<----------TRAV REGION----------> (Proximal V Genes (3' end) … Distal V Genes (5' end))

PCR results Max Cycle-Diff G3PDH		TRAJ61	TRAJ56	TRAJ53	TRAJ48	TRAJ47	TRAJ41	TRAJ33	TRAJ24	TRAJ10	TRAJ5
TRAV41	Sample 1	6,20 Average	NS Average	7,30 Average	7,50 Average	6,00 Average	NS Average	6,30 Average	3,30 Average	0,00 Average	0,00 Average
	Sample 2	6,30 6,60	NS 3,00	7,35 7,13	6,30 7,33	4,95 5,82	NS	6,90 6,34	2,55 3,08	1,75 0,44	0,00 0,00
	Sample 3	5,25 St. Dev.	1,50 St. Dev.	6,30 St. Dev.	7,50 St. Dev.	NS St. Dev.	NS St. Dev.	5,00 St. Dev.	0,00 St. Dev.	0,00 St. Dev.	0,00 St. Dev.
	Sample 4	8,65 1,45	4,50 2,12	7,55 0,56	8,00 0,72	6,50 0,79	NS	7,15 0,96	6,45 2,66	0,00 0,88	0,00 0,00
TRAV40	Sample 1	7,75 Average	6,30 Average	7,40 Average	7,65 Average	7,60 Average	NS Average	6,65 Average	3,15 Average	1,95 Average	0,00 Average
	Sample 2	7,45 7,25	6,10 5,81	6,65 6,51	8,25 7,83	7,30 7,04	5,35 5,78	5,95 5,80	4,35 4,86	0,00 1,11	0,00 0,00
	Sample 3	5,40 St. Dev.	4,30 St. Dev.	5,00 St. Dev.	7,40 St. Dev.	5,25 St. Dev.	NS St. Dev.	4,30 St. Dev.	5,45 St. Dev.	2,50 St. Dev.	0,00 St. Dev.
	Sample 4	8,40 1,30	6,55 1,03	7,00 1,05	8,00 0,38	8,00 1,23	6,20 0,60	6,30 1,04	6,50 1,44	0,00 1,30	0,00 0,00
TRAV38	Sample 1	8,00 Average	9,50 Average	9,00 Average	10,00 Average	8,50 Average	8,35 Average	8,00 Average	6,50 Average	7,00 Average	4,00 Average
	Sample 2	7,40 7,18	NS 8,75	8,00 8,33	8,00 8,25	7,00 7,53	NS 7,45	7,00 7,08	5,00 6,13	5,80 6,33	0,00 1,50
	Sample 3	6,00 St. Dev.	8,00 St. Dev.	8,00 St. Dev.	8,00 St. Dev.	7,10 St. Dev.	7,00 St. Dev.	6,00 St. Dev.	6,00 St. Dev.	7,00 St. Dev.	2,00 St. Dev.
	Sample 4	7,30 0,84	9,00 0,76	NS 0,58	7,00 1,26	7,50 0,68	7,00 0,78	7,30 0,83	7,00 0,85	5,50 0,79	0,00 1,91
TRAV35	Sample 1	5,10 Average	6,45 Average	8,70 Average	NS Average	NS Average	6,75 Average	6,48 Average	6,55 Average	4,65 Average	0,00 Average
	Sample 2	4,50 5,65	5,90 6,34	8,10 8,05	6,35 6,78	NS 6,38	6,15 6,30	5,85 5,82	4,90 6,78	3,35 4,25	0,00 0,11
	Sample 3	6,00 St. Dev.	6,00 St. Dev.	7,00 St. Dev.	7,00 St. Dev.	5,40 St. Dev.	6,00 St. Dev.	4,50 St. Dev.	8,25 St. Dev.	4,00 St. Dev.	0,00 St. Dev.
	Sample 4	7,00 1,09	7,00 0,50	8,40 0,74	7,00 0,38	7,35 1,38	6,30 0,32	6,45 0,93	7,40 1,43	5,00 0,73	0,45 0,23
TRAV30	Sample 1	1,85 Average	4,75 Average	6,40 Average	NS Average	NS Average	4,40 Average	5,10 Average	5,80 Average	5,40 Average	2,75 Average
	Sample 2	1,15 3,13	4,15 4,48	5,20 5,84	6,70 5,70	6,25 7,08	3,05 4,64	3,60 4,90	5,05 5,43	NS 7,13	0,70 3,86
	Sample 3	4,00 St. Dev.	3,00 St. Dev.	5,50 St. Dev.	5,40 St. Dev.	10,00 St. Dev.	5,50 St. Dev.	5,50 St. Dev.	9,40 St. Dev.	11,00 St. Dev.	6,00 St. Dev.
	Sample 4	5,50 1,99	6,00 1,25	6,25 0,58	5,00 0,89	5,00 2,60	5,60 1,19	5,40 0,88	5,00 0,53	5,00 3,35	6,00 2,61
TRAV21	Sample 1	6,60 Average	7,20 Average	7,00 Average	NS Average	NS Average	6,40 Average	6,60 Average	6,10 Average	5,90 Average	5,10 Average
	Sample 2	6,20 5,79	6,90 6,53	6,95 6,36	6,10 5,37	5,20 5,40	6,15 5,75	5,80 5,44	5,40 5,21	5,05 5,61	3,65 4,81
	Sample 3	5,00 St. Dev.	7,00 St. Dev.	6,50 St. Dev.	5,00 St. Dev.	5,00 St. Dev.	6,00 St. Dev.	5,00 St. Dev.	5,00 St. Dev.	5,00 St. Dev.	4,00 St. Dev.
	Sample 4	5,35 0,74	5,00 1,02	5,00 0,94	5,00 0,64	6,00 0,53	4,45 0,88	4,35 0,98	4,35 0,73	6,50 0,72	6,50 1,28
TRAV16	Sample 1	4,80 Average	6,20 Average	7,10 Average	NS Average	NS Average	4,30 Average	5,50 Average	6,20 Average	6,90 Average	3,20 Average
	Sample 2	6,35 4,93	7,80 6,00	8,50 7,01	9,10 6,37	7,40 6,70	NS 4,27	7,05 5,98	7,35 7,00	8,10 8,09	3,50 3,30
	Sample 3	4,00 St. Dev.	4,50 St. Dev.	6,00 St. Dev.	5,00 St. Dev.	6,00 St. Dev.	3,50 St. Dev.	NS St. Dev.	7,00 St. Dev.	8,00 St. Dev.	3,00 St. Dev.
	Sample 4	4,55 1,01	5,50 1,39	6,45 1,09	5,00 2,37	5,00 0,99	5,00 0,75	5,40 0,93	7,45 0,57	9,35 1,00	3,50 0,24
TRAV2	Sample 1	0,00 Average	0,00 Average	0,00 Average	0,00 Average	0,00 Average	3,15 Average	3,75 Average	6,45 Average	3,10 Average	4,05 Average
	Sample 2	0,00 0,00	0,00 1,00	0,00 0,70	0,00 0,75	1,80 3,08	NS 3,55	NS 4,92	7,15 7,69	3,80 6,28	5,80 5,91
	Sample 3	0,00 St. Dev.	3,00 St. Dev.	2,00 St. Dev.	2,00 St. Dev.	8,50 St. Dev.	4,00 St. Dev.	4,00 St. Dev.	9,25 St. Dev.	10,40 St. Dev.	6,00 St. Dev.
	Sample 4	0,00 0,00	1,00 1,41	0,79 0,94	1,00 0,96	2,00 3,73	3,50 0,43	7,00 1,81	7,90 1,20	7,80 1,84	7,80 1,53
Sums of the average values		41,64	43,03	50,56	49,37	50,01	39,63	47,56	53,38	47,62	25,50
		Standart deviation max:		3,73		Average standart deviation:		1,06		Total averages:	418,62

LEGENDS :	Color code:	0 to 4	4 to 5	5 to 6	6 to 7	up to 7	0,00 : no detected	NS: presence of Non Specific band(s) after gel migration of PCR products

Figure 2. Real time PCR quantifications of 80 V-J associations form 4 human thymic samples. Results are expressed in arbitrary units indicating the difference in cycle number at which products were first detected. Amplifications of the G3DPH gene were used as normalizer. Quantifications were performed from 4 thymic genomic DNA samples corresponding to 4 children aged between 1 month and 1 year old.

at both allelic loci for each simulated cell and, as a reminder, the entire values of parameters corresponding to the simulation.

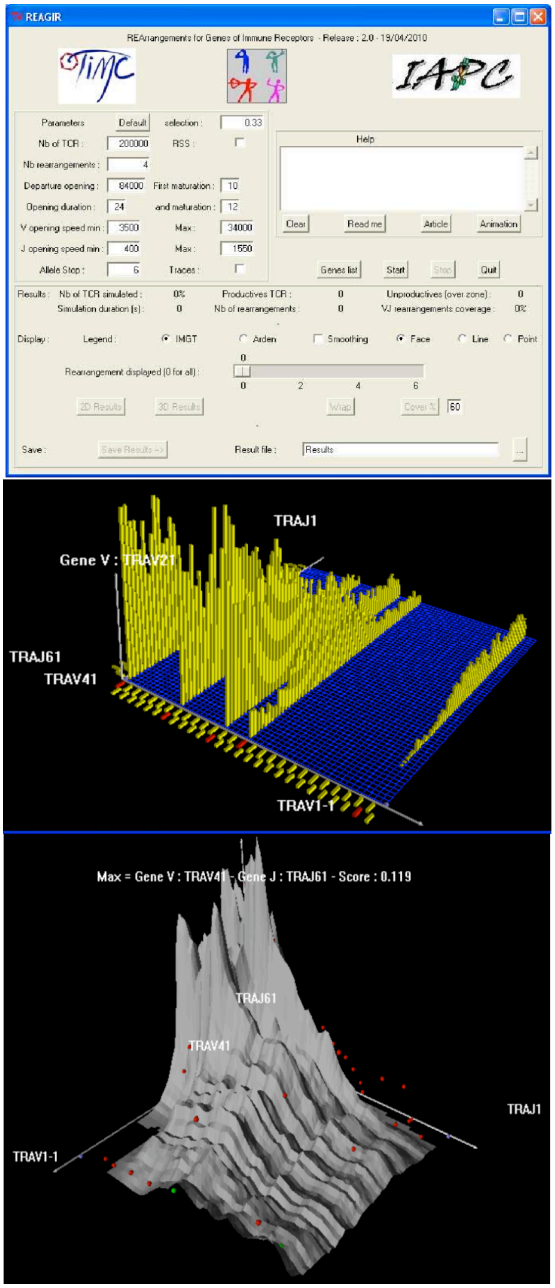


Figure 3. Program interface and generated graphs. On the graph above, V genes displaying their J region use are interactively set by the user.

Parameter values, robustness of the human model, and comparison with a prior study conducted on mouse

The stochastic successive windowing model was firstly developed for the Vα-Jα use during mouse TRA locus successive rearrangements. For this species, the ontogeny days when V and J genes were first seen rearranged in conjunction with physical gene positions and gave speeds of progression for the gene accessibility to rearrangements (Pasqual, 2002; Thuderoz, 2010). These speeds, denoted V and J region opening speeds, were successfully used in mouse model as simulation results and fitted thymic quantifications from experiments only by using two opening speeds chosen within intervals closed to the experimental 99.9% confidence intervals. In fact, the V speed (denoted S_V) belonged to the speed interval [0.35 Kb/h, 34 Kb/h] and the J speed (S_J) to [0.4 Kb/h, 1.55 Kb/h] with a mean opening speed of roughly 18 Kb/h for the V region and 1 Kb/h for the J region. In the transition to human, structural specificities of the human TRA/TRD locus were integrated to the model. Concerning opening speeds, values of the J region opening speed interval determined from experience in mouse were used in the human model in order to be consistent with the high conservation of both the J region structure and the J gene sequences between the two species (Uenishi *et al*, 2003). On the other hand, the V opening speed interval was determined by simulations: interval [0.5 Kb/h, 15.65 Kb/h] (average 8 Kb/h) provided simulation results that fitted the best human experimental distributions (least squares method). The opening location of the simulation was fixed between the V and J genes in order to access directly to the TRAV and TRAJ genes after the first maturation, which was set to allow the elimination of the TRD genes in the model (Fig. 1). For humans, as well as in mice, issues obtained from the modeling that fit the best the experimental data, indicated the duration of the first maturation step had a mean value of 5 hours and the opening duration before each rearrangement was 24 hours. Interestingly, parametric study gave distinct numbers of successive rearrangements between the two species

with 3-4 for mouse and 4-5 for human. To make certain the sampling size used in simulations was large enough, the representativeness of the repertoire was tested by making sets of simulations of increasing size. Diversity became constant when the population size was higher than 5×10^5 T cells, showing the pertinence of a repertoire calculation based on a 10^6 alpha chain population (Arstila *et al*, 1999). Tests of robustness performed on the model indicated that variations of about 5 to 10% in the values of the parameters provided simulation results statistically coherent with experimental data, though larger variations induced major deviations on the modeling simulation results inconsistent with (i.e. significantly different from) experimental data.

Stochastic successive windowing model results fit experimental V-J quantifications from human thymi

In order to compare simulation results to experiments, the very 80 V-J associations experimentally examined were extracted from the simulated results computed for the totality of the V-J association frequencies in the framework of the successive windowing model. Results from simulations were transformed in the base of the 90% of amplification efficacy of real time PCRs ($\log_{1,8}$(simulated frequency)), normalized over the total sum of PCR average results. Figure 4 showed simulated results generated using the set of parameters fully described above; the Figure indicates that results obtained with 4 and 5 successive rearrangements (R=4 and R=5) for the two simulated populations fit experimental results. Numbers in bold indicated that simulated values lie within the two standard deviations confidence interval of the experimental PCR quantifications, corresponding to the fact that the alpha risk of rejection of the hypothesis of a difference between predicted and observed levels of gene rearrangements is more than 0.25. At last, the consistency between simulated data and frequencies determined from thymic genomic DNA validates our model as a relevant tool accounting for the dynamical building of the TRα combinatorial repertoire in human thymus.

Successive windowing model results fit interallelic distance for J segment use

Simulation results were tested in parallel for their coherence with V-J quantifications and for their ability to fit an interallelic J distance distribution issued from the

Simulation results generated with R=4

genes	TRAJ61	TRAJ56	TRAJ53	TRAJ48	TRAJ47	TRAJ41	TRAJ33	TRAJ24	TRAJ10	TRAJ5
TRAV41	**7,13**	7,27	**7,20**	**7,13**	**6,82**	0,00	1,95	**0,00**	**0,00**	**0,00**
TRAV40	**7,00**	**6,98**	**7,09**	6,93	**6,31**	**5,84**	3,19	0,00	**0,00**	**0,00**
TRAV38	**7,37**	**7,33**	**7,48**	**7,44**	**6,99**	**6,34**	4,45	3,06	0,00	**0,00**
TRAV35	**7,53**	7,57	**7,60**	7,71	**7,45**	**6,99**	**6,51**	**5,56**	**4,63**	3,51
TRAV30	**7,07**	7,10	7,31	**7,34**	**7,18**	**6,81**	**6,40**	**5,75**	**4,68**	**4,61**
TRAV21	**7,23**	**7,16**	**7,14**	7,58	7,22	**6,67**	**6,52**	**6,18**	**5,64**	**5,53**
TRAV16	**5,74**	**5,33**	**5,65**	**6,34**	**6,44**	6,08	**5,96**	5,48	5,00	5,03
TRAV2	**0,00**	**0,00**	**0,00**	3,73	**3,73**	**3,97**	**5,60**	**5,51**	**5,79**	**5,74**
Sums:	49,09	48,75	49,48	54,19	52,13	42,70	40,59	31,53	25,74	24,42
									Total:	418,62

Simulation results generated with R=5

genes	TRAJ61	TRAJ56	TRAJ53	TRAJ48	TRAJ47	TRAJ41	TRAJ33	TRAJ24	TRAJ10	TRAJ5
TRAV41	**7,32**	7,28	**7,19**	**7,09**	**6,76**	0,00	1,92	**0,00**	**0,00**	**0,00**
TRAV40	**7,10**	**7,04**	**7,06**	6,86	**6,28**	**5,80**	3,13	0,00	**0,00**	**0,00**
TRAV38	**7,45**	**7,37**	**7,47**	**7,40**	**6,99**	**6,34**	4,70	3,00	0,00	**0,00**
TRAV35	**7,57**	7,54	**7,54**	7,69	**7,43**	7,03	**6,57**	**5,65**	**4,85**	3,75
TRAV30	**7,02**	7,00	7,21	**7,28**	**7,18**	**6,84**	**6,47**	**5,87**	**5,00**	**4,90**
TRAV21	**7,10**	**7,03**	**7,01**	7,47	7,11	**6,61**	**6,54**	**6,14**	**5,60**	**5,59**
TRAV16	**5,64**	**5,23**	**5,55**	**6,24**	**6,32**	6,00	**5,94**	5,59	5,20	5,09
TRAV2	**0,00**	**0,00**	**0,00**	3,66	**3,66**	**3,89**	**5,50**	**5,40**	**5,81**	**5,74**
Sums:	49,20	48,49	49,03	53,70	51,72	42,51	40,78	31,66	26,45	25,08
									Total:	418,62

Color code:	0 to 4	4 to 5	5 to 6	6 to 7	up to 7

Figure 4. V-J association frequencies $F_{Vi\text{-}Jj}$ from simulations converted in theoretical PCR cycle numbers ($\log_{1.8}(F_{Vi\text{-}Jj})$). Parameter values used to generate these simulated results are fully presented in the Results section. The Figure indicates that results obtained with 4 and 5 successive rearrangements (R=4 and R=5) for the two simulated populations fit experimental results. Numbers in bold indicate that simulated values lie within the two standard deviation confidence interval of experimental PCR quantifications, corresponding to the fact that the alpha risk of rejection of the hypothesis of a difference between predicted and observed levels of gene rearrangements is more than 0.25.

literature (Davodeau *et al*, 2001). The involvement of a parameter denoted $Stop_{delay}$ within the model allowed computing the interallelic behavior at the very moment that TR-mediated signals stopped successive rearrangements in a given simulated cell (see M&M section). Figure 5 shows that fixing $Stop_{delay}=0$ for all simulated cells gave an interallelic usage of J genes too close compared to experimental distribution (Fig. 5.A). On the contrary, a systematic stop of successive rearrangements within the subsequent rearrangement round ($Stop_{delay}=1$) gave an interallelic usage of J genes, which are too distant compared to the experimental distribution (Fig. 5.D). Simulations using $Stop_{delay}=0$ and $Stop_{delay}=1$ in proportions 25%/75% or 75%/25% of the cells gave two distributions that fit the experimental distribution (Fig. 5.B and 5.C). Beyond a satisfactory result concerning successive windowing model validation, the consequences of this fit regarding the putative interallelic behavior during the stop of the successive rearrangements were addressed in the Discussion section. Briefly, to complete these results, an identical analysis was conducted on mouse (additional Fig. 1). In this species, four studies addressing Jα usage at both TRA alleles available in the literature were compiled in order to get a more significant database (n=110), giving an empirical distribution of the interallelic distances in J usage statistically more confident than in the human case (n=29) (Davodeau *et al*, 2001; Heath *et al*, 1995; Malissen *et al*, 1992; Casanova *et al*, 1991). The same conclusions emerge in mouse as in human, rejecting simulations performed using $Stop_{delay}=0$ or $Stop_{delay}=1$ for the whole cells (additional Fig. 1.A and 1.D). However, the more precise experimental distribution allows stating between the B and D graphs: simulations using $Stop_{delay}=0$ and $Stop_{delay}=1$ in the respective proportions of 75% and 25% fit the best the experimental distribution (additional Fig. 1.B)

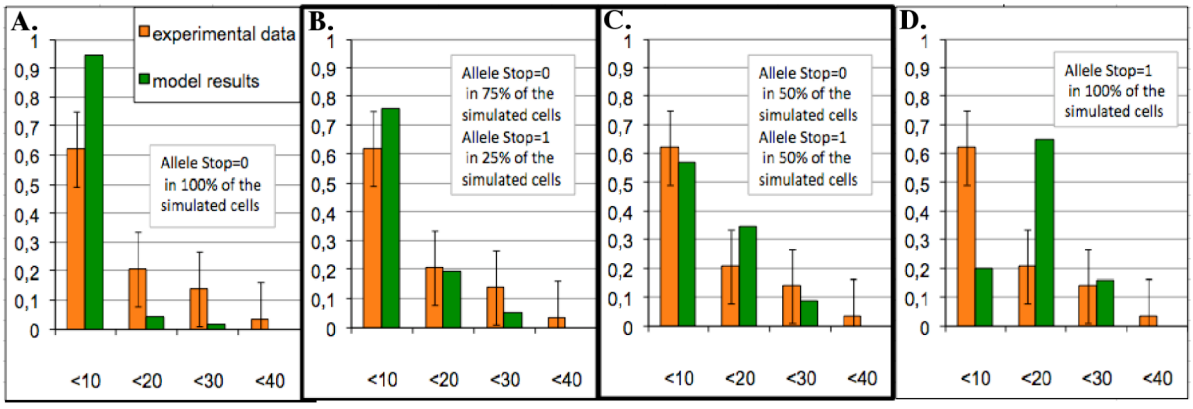


Figure 5. Experimental J interallelic distance distribution presented along with model results generated using different values for the $Stop_{delay}$ parameter. The experimental distribution reproduced on the four graphs was taken from Figure 2 in (Davodeau, 2001), which presented the distribution of the differences in rank of Jα segments used at both alleles in several clones (authors denoted these values interallelic J distances). Error bars were added on the distribution for n=29; the formula *(freq.*(1-freq.))/n* was used for calculating each frequency (*freq.*) variance in histogram classes. The four graphs compare the experimental distribution with parametric study results generated varying the $Stop_{delay}$ parameter value (described in M&M). Comparison between experimental distribution and model simulation results showed that a stop of successive rearrangements within the same rearrangement round ($Stop_{delay}=0$) gave an interallelic usage of J genes too close compared to experimental distribution (A.). On the contrary, a systematic stop of successive rearrangements within the subsequent rearrangement round ($Stop_{delay}=1$) gave interallelic usage of J genes too distant compared to the experimental distribution (D). Finally, parametric studies predict that when an in-frame rearrangement is successfully selected in a cell, the intracellular signaling will stop further rearrangements over the other allele during the same rearrangement round for 50% to 75% of the cells and during the subsequent rearrangement round for the rest of the cells (B. or C).

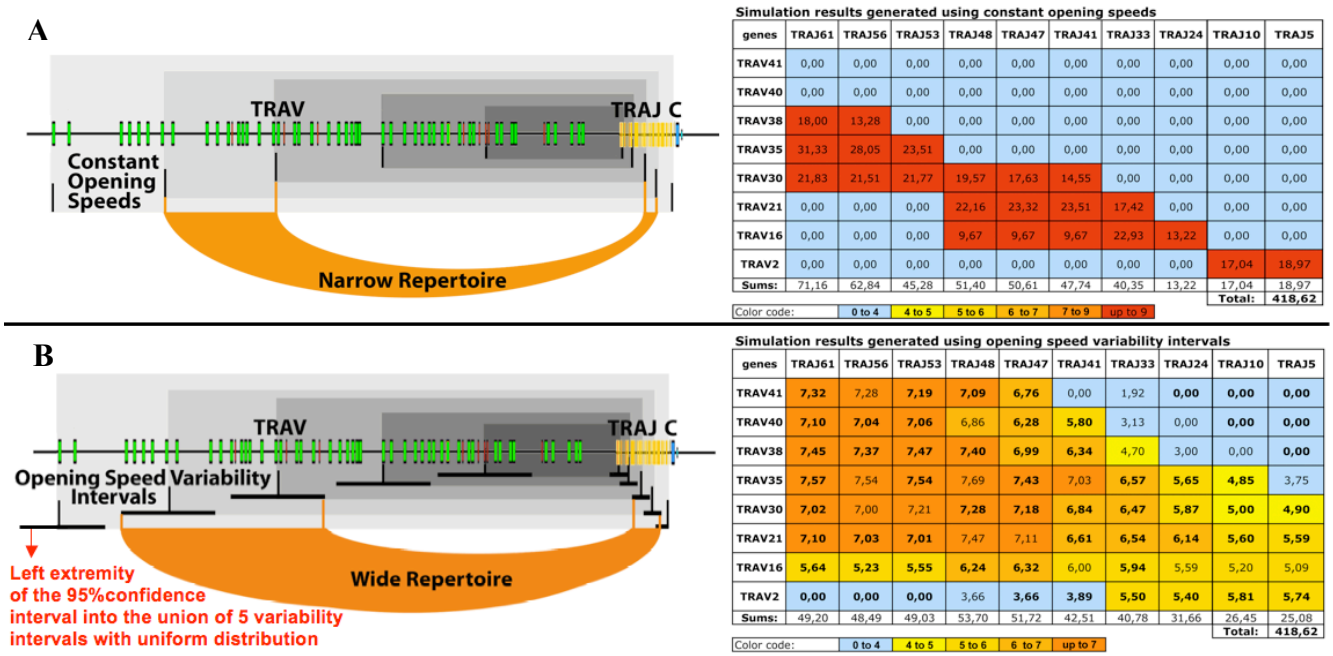
Discussion

Human thymic TRα combinatorial repertoire from extensive real time PCR analyses

In spite of the human TRA gene polymorphism and the use of an elevated number of primers within the same PCR amplifications, the 80 V-J quantifications from thymic genomic DNA displayed a good inter-individual consistence among the 4 samples tested (maximal standard deviation over the entire experiments: 3.73; average standard deviation: 1.06). Quantitative PCR analysis results (Fig. 2) undoubtedly demonstrated each V gene tested used a subsequent subset of J genes according to their position over the locus, as we previously observed on a fewer number of V and J genes, using multiplex PCR analyses (Jouvin-Marche *et al*, 2009). It is worth to notice that the outside parts of these subsets formed two areas of non-observed or infrequent associations, corresponding to distal V / proximal J and proximal V / distal J (appearing in blue on Fig. 2) and that within each subset, frequencies differ highly. Finally, the corresponding 8 Vα and 10 Jα genes investigated, chosen regularly spread along the Vα and Jα regions, offered a first experimental wide-ranging sampling of the human TRα computational repertoire in the thymus.

Explanatory power of inside out gene use

If the inside out use of the genes was presumed to cause unequal frequencies of V-J gene associations (Krangel, 2009), the successive windowing model definitely demonstrated that a precise progression of gene accessibility to rearrangements together with an opening of the J region synchronized between both TRA alleles were sufficient to fully explain the experimental V-J frequencies currently available as well as the experimental interallelic J usage. Given the prominent conservation of J region structure between mouse and human species, the accessibility progression over the J region experimentally determined from ontogeny analyses in mice was used for the human model. This J region opening speed consisted in a variable speed (S_J) belonging to the [0.4 Kb/h, 1.55 Kb/h] variability interval, corresponding to a mean opening speed of about 1 Kb/h. Parametric studies performed for human models showed that this experimental J speed variability interval was the unique one allowing the generation of V-J association profiles fitting experimental data (a bias superior of 10% of the mean speed generated data incoherent with biological results), as it was already demonstrated for mouse (Thuderoz *et al*, 2010). Hence, model studies, performed for both the human and mouse species, predicted J regions, highly conserved in terms of genes number, length, and regulatory Cis elements, would proceed in a similar progression of accessibility through the successive rearrangement process. Concerning the V region accessibility progression, experiments performed during mouse ontogeny allowed the calculation of a mean speed S_{Vmouse}= 18Kb/h. Mouse model showed also the necessity to introduce a variability interval for the opening V speed equal to [0.35 Kb/h, 34 Kb/h] and centered on this mean value, in order to get distributions in accordance with experimental data (with an acceptable bias of 10% on the interval). Regarding the human V opening speed, major differences between mouse and human V regions avoided making preliminary assumptions and we have retained the opening speed variability interval [0.5 kb/h, 15.650 kb/h] after extensive parametric studies. The models developed for both species predict that the recombination centers (constituted over the J region throughout the successive V-J recombinations process) and human V regions.



Simulation results generated using constant opening speeds

genes	TRAJ61	TRAJ56	TRAJ53	TRAJ48	TRAJ47	TRAJ41	TRAJ33	TRAJ24	TRAJ10	TRAJ5
TRAV41	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRAV40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRAV38	18,00	13,28	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRAV35	31,33	28,05	23,51	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRAV30	21,83	21,51	21,77	19,57	17,63	14,55	0,00	0,00	0,00	0,00
TRAV21	0,00	0,00	0,00	22,16	23,32	23,51	17,42	0,00	0,00	0,00
TRAV16	0,00	0,00	0,00	9,67	9,67	9,67	22,93	13,22	0,00	0,00
TRAV2	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	17,04	18,97
Sums:	71,16	62,84	45,28	51,40	50,61	47,74	40,35	13,22	17,04	18,97
									Total:	418,62

Color code: 0 to 4 | 4 to 5 | 5 to 6 | 6 to 7 | 7 to 9 | up to 9

Simulation results generated using opening speed variability intervals

genes	TRAJ61	TRAJ56	TRAJ53	TRAJ48	TRAJ47	TRAJ41	TRAJ33	TRAJ24	TRAJ10	TRAJ5
TRAV41	7,32	7,28	7,19	7,09	6,76	0,00	1,92	0,00	0,00	0,00
TRAV40	7,10	7,04	7,06	6,86	6,28	5,80	3,13	0,00	0,00	0,00
TRAV38	7,45	7,37	7,47	7,40	6,99	6,34	4,70	3,00	0,00	0,00
TRAV35	7,57	7,54	7,54	7,69	7,43	7,03	6,57	5,65	4,85	3,75
TRAV30	7,02	7,00	7,21	7,28	7,18	6,84	6,47	5,87	5,00	4,90
TRAV21	7,10	7,03	7,01	7,47	7,11	6,61	6,54	6,14	5,60	5,59
TRAV16	5,64	5,23	5,55	6,24	6,32	6,00	5,94	5,59	5,20	5,09
TRAV2	0,00	0,00	0,00	3,66	3,66	3,89	5,50	5,40	5,81	5,74
Sums:	49,20	48,49	49,03	53,70	51,72	42,51	40,78	31,66	26,45	25,08
									Total:	418,62

Color code: 0 to 4 | 4 to 5 | 5 to 6 | 6 to 7 | up to 7

Figure 6. Importance of the variability of the V and J region opening speeds in the building of the combinatorial repertoire diversity. Schemes represent the TRA locus after deletion of the TRD locus (V genes appear in green, J genes in yellow and the single C gene is shown in blue). The occurrence of 5 accessibility windows (gray rectangles) corresponds to the 5 rearrangement rounds. **A.** The use of constant opening speeds in simulations generates a narrow repertoire, as shown in simulation results (top right). **B.** Variable opening speeds for the V and the J regions appear to constitute an important factor for generating TRα repertoire diversity. The wide repertoire (bottom right) was simulated using opening speed variability intervals.

These distinct opening speed variability intervals, specific of the two species, seem to be proportional to the V region length, hence allowing to take advantage of the diversity offered by the totality of the V genes, independently of the length of the V region. In addition, simulations performed within the distinctive frameworks of human and mouse models showed the use of speeds variable inside an interval was important for the diversity of the repertoire generated. In fact, the use of constant opening speeds in simulations gave a narrower use of V-J associations, increasing considerably the two zones of improbable associations observed from experiments (depicted in blue cells on Fig. 2). A scheme presented in Figure 6 illustrates, for each rearrangement round, the place of the accessibility windows over the TRA locus (gray rectangles). On Figure 6.A, the use of constant opening speeds for the V and J regions are shown to provoke well-segmented accessibility windows throughout the simulation, which resulted in the building of a narrow repertoire, with few hyper-represented V-J associations (shown in red cells). On the other hand, the Figure 6.B illustrated how the empirical opening speed variability intervals for V and J regions accessibility progression generated accessibility windows carrying a variability in terms of size and position, causing a wider repertoire, coherent with experiments. Human model shows that progression of gene accessibility to rearrangements over TRAJ and TRAV regions (with human V speed different from the mouse one) constitutes a mechanism sufficient to explain the relative abundances of the V-J association frequencies determined from real time PCR assays on human thymic genomic DNA.

Stop of successive rearrangements: putative interallelic behavior

Regarding the J region, it is well accepted both J alleles perform a synchronized opening, which provokes an interallelic use of J segments separated by less than 10 genes for the majority of the lymphocytes (Villey *et al*, 1996). Nevertheless, minor occurrences of more distant interallelic J genes use are

observed in experiments (Davodeau *et al*, 2001). When a Vα-Jα in-frame rearrangement is performed over an allele, beta pairing, surface expression and positive selection are all needed to allow the TR-mediated signals stopping further rearrangements in the cell. As commonly accepted, the feedback loop due to positive selection of a TRα chain may logically inhibit further rearrangements on the very TRA allele that originated its synthesis. Concerning the other allele, simulations showed a stop of successive rearrangements within the same rearrangement round ($Stop_{delay}$=0) gave interallelic usage of J genes too close compared to experimental distribution (Fig. 5.A). On the contrary, a systematic stop of successive rearrangements within the subsequent rearrangement round ($Stop_{delay}$=1) gave interallelic usage of J genes too distant compared to the experimental distribution (Fig. 5.D). Hence, the involvement of the Allele Stop parameter inside the model suggests a putative scenario for the other allele behavior: in a majority of cells (supposedly from 50% to 75% of the cells, Fig. 5.B and 5.C), the feedback loop would inhibit further rearrangements on this other allele within the same rearrangement round, but in the rest of the cells, a slight out of phase between two allele rearrangements would allow the other allele to perform an extra rearrangement round before TR-mediated signals step in and stop the mechanism. To complete this analysis, additional results in mouse, based on a larger experimental data base proposed the feedback loop would inhibit further rearrangements on the other allele during the same rearrangement round in the clear majority of the cells (75%, as on the additional Fig. 1.B), the extra rearrangement being performed only in a minority of cells (25%).

RSS: refining local frequencies on the repertoire shape

The human model included Recombination Signal Sequence (RSS) diversity effect through RSS scores, which were calculated from the percentage of homology of each RSS sequence with the consensus.

The facultative use of RSS scores in simulations allowed observing their influence on results. RSS sequences appeared to change not the global repertoire shape, but only local specificities. This is in good accordance with mouse TRB locus observations that demonstrated V gene RSSs neither correlate with any specific restriction in use of the subset of J genes nor with any elevated V-J rearrangement frequencies (Wilson *et al*, 2001).

Eventually, the RSSs would hold less impact in terms of gene use frequencies than in bi-directional use of the TRA genes during the successive rearrangements process.

Conclusions for the human TRα repertoire size

The potential TRα repertoire was first estimated to $0.5*10^6$ chains in human blood through CDR3 heterogeneity analysis and considering that every V-J association was obtained from independent events consisting in tossing a couple of genes V and J with their marginal frequency (Arstila *et al*, 1999). In this estimation, every one of the 54 V genes was supposed to rearrange every one of the 61 J genes, giving a combinatorial diversity of 3294 associations. Identification of non-functional genes first limited this evaluation. In a previous work, experimental quantifications performed from blood and thymic material stated any one of the three V genes tested (V41, V40 and V1) rearranged not the entire J region but a subset of subsequent J genes, according to the V position in the locus. Consequently, combinations outside every

J subsets, corresponding to non-observed or rare V-J associations, were presumed to lower the potential TRα combinatorial repertoire and a preliminary estimation of approximately 2000 associations in periphery and thymus was announced (Jouvin-Marche *et al*, 2009). In the present study, the J region use was tested through the use of ten J genes by eight V genes, using genomic DNA quantifications from human thymic samples. This further comprehensive experimental study defines more precisely the limits of the non-observed or rare V-J association areas (appearing in blue cells on Fig. 2). At a mechanistic level, the successive windowing model proposed a dynamical explanation for the occurrence of the proximal V / distal J and distal V / proximal J genes rare associations: they would originate from a non-synchronized placement. More, this stochastic model, validated for the human thymic V-J quantifications, allows calculating the V-J frequencies for the entire V-J associations. Knowledge about these frequencies allows proposing a new update on the potential TRα combinatorial repertoire evaluation. In fact, if about 2100 V-J associations would correspond to the totality of the possible combinations, some of these associations are considerably more abundant than others. Table 1 indicates the size of the most abundant V-J associations along with their frequency. If roughly 1700 associations would correspond to 95% of the α chains, about 1000 associations would correspond to 75%, and approximately, the 500 most abundant V-J associations would represent a half of the the total. These observations point out the α chain combinatorial repertoire diversity supporting immuno-competence was highly over-estimated in thymic cells: the diversity may consist essentially in roughly 500 strongly represented V-J associations accounting for about the half of the V-J associations found in α chains, along with 1500 other less represented V-J associations.

Number of the most abundant V-J associations	Percentage of the α chains corresponding to the most abundant V-J associations
2125	100%
1721	95%
954	75%
453	50%

Table 1. Number of the most abundant V-J associations along with the percentage of the α chains they represent. Table data were compiled using the entire V-J associations frequencies computed from the human successive windowing model.

Number of rearrangement rounds expected in human TRA loci

More than offering information on the α repertoire shape and diversity, the combined experimental and model approaches allow making predictions on the values of intrinsic parameters not accessible through experience like the very number of rearrangement rounds performed through the TRA rearrangement mechanism. Whereas the mouse model predicted a total of 3 to 4 successive rearrangements performed in mouse TRA locus (Thuderoz *et al*, 2010), in human however, the model parametric studies gave an estimation of 4 to 5 rearrangement rounds. Simulations showing the occurrence of four or five rearrangements would favor the use of the distal genes of V and J regions in human. In fact, the use of a fewer number of rearrangements in simulations gave an under-representation of distal J / distal V associations. At last, the incidence of a supplementary rearrangement would give the human species one additional chance per allele to generate an in-frame rearrangement, which may constitute a major selective advantage compared to the mouse, compensating for the lower number of human V genes compared to that in mouse.

Conclusion

The stochastic successive windowing model presented throughout this paper

stands that the precise progression of gene accessibility to rearrangements according to non-constant opening speeds along with an interallelic synchronized J opening within each cell constitutes a sufficient mechanism to explain experimental V-J frequencies and interallelic J usage. Eventually, the modeling step, using a multi-level systemic approach followed by a simulation phase, offered a clear understanding of the dynamics building of the human α repertoire, in order to propose predictions on this repertoire diversity richness and to dispose of a simulated theoretical repertoire showing the frequencies of the entire V-J associations. Knowledge about the human thymic repertoire shape (Figure 3) constitutes indeed a key issue in the immune system development, and thus a crucial requirement in therapeutic interventions aiming to reconstitute or to control immune responses.

Material and Methods

Experimental approach: quantification of Vα-Jα associations from human thymus extracts

Gene Nomenclature

Gene names correspond to the ImMunoGeneTics (IMGT) nomenclature (http://imgt.cines.fr). NCBI (National Center for Biotechnology Information) accession numbers are AE000658-AE000662 for the V region and M94081 for the J region. Positions of the V and J genes in the TRA/TRD locus were determined from the first nucleotide of the TRAC gene as previously described (Baum *et al*, 2004).

Provenance of the samples

The experimental dataset presented in this issue was generated with genomic DNA extracted from thymi of 4 healthy children, between 1 month and 1 year old. The thymi were from surgical waste and the committee approving the experiments, and checking that informed consent was obtained from all subjects, has been the same than those noticed in the previously published paper (Fuschiotti *et al*, 2007).

Selection and design of primers

The Vα and Jα primers were selected in order to globally cover the V and J regions and were spaced regularly between 50 to 60 kb and 5 to 10 kb for the V and J regions respectively. The specificity of each primer was checked using the BLAST sequence alignment program (Altschul *et al*, 2009). The selection criteria for the primers were:

1. The Recombination Signal Sequence (RSS) of the V or J genes were quite identical to the consensus sequence.
1) The optimal amplification temperatures were in the same range of order for all the primers, allowing the use of identical amplification conditions for all the PCR.
2) The amplification yield of each couple of primer was 90%, thus allowing a direct relative comparison.
3) Each couple of primers gave only one band on agarose gel and size of each amplification product was similar, allowing comparing PCR results.

The list of the primers designed is presented on Table 2. would recruit progressively V genes causing a differential speed of use between mouse

PCR conditions

Real time PCRs were performed on a Light Cycler™ (Roche Diagnostics, Meylan, France). According to the primer features, the quantitative PCR were all realized in the following conditions: Denaturation 95°C, 10 min.; Amplification (40 cycles); Denaturation 94°C, 15 sec.; Hybridation 67°C, 7sec. ; Elongation 72°C, 7 sec.

Primer	5'-3' Sequence
hTRAV2	TCTCTTCATCGCTGCTCATCCTCC
hTRAV 16	AGAGTGACTCAGCCCGAGAAG
hTRAV 21	TGCCTCGCTGGATAAATCATCAGG
hTRAV 30	GCCGTGATCCTCCGAGAAGGGG
hTRAV35	GGCTGGGAAGTTTGGTGATATAGTGTC
hTRAV38	AGCAGCCAAATCCTTCAGTCTCAA
hTRAV40	AAGACAAAAACTCCCCCATTGTGAAATA
hTRAV41	GCCCTCCTGAAAATGTGTAAAGAAATGT
hTRAJ5	CTGTCTCTGCAATGATGAAATGGCC
hTRAJ10	CCACTTTTAGCTGAGTGCCTGTCCC
hTRAJ24	GGTCCCTGCTCCAAACTGC
hTRAJ33	CGCCCCAGATTAACTGATAGTTGCT
hTRAJ41	TGCCCCGAGACCTGATAACCAA
hTRAJ47	GGGTTGCCTTCGAGAGCGTTAATC
hTRAJ48	AGCACTTGACGGCAGCAGC
hTRAJ53	CTTCCCCCACTCCCTTCAAACTTAC
hTRAJ56	ACTGGGCAGGAGATTCGGTTAT
hTRAJ61	ACTTGCTGAGTTTCATGATTCCTC
G3DPHup	AGCAATGCCTCCTGCACCACCAAC
G3DPHdo	CCGGAGGGGCCATCCACAGTCT

Table 2. Primers designed in order to perform quantitative PCRs with genomic DNA extracted from human thymi. The ImMunoGeneTics nomenclature was used for the V and J genes (http://imgt.cines.fr/). Accession numbers: AE000658-AE000662 for the V region; M94081 for the J region.

Standardization of the experimental quantifications

A number of 80 V-J combinations were quantified for each of the 4 human thymic samples among a total of 2250-2350 potential associations (considering that 45 to 47 V genes and 50 J genes are functional). The quantifications of the V-J associations (Fig.4) were standardized among the PCR experiences according to the house keeping genes quantification values, used as a normalizer:

$$\text{V-J quantif.} = \text{Ct(house keeping gene amplification product)} - \text{Ct(V-J association)},$$

where Ct is the Cycle threshold, cycle at which the sample reaches the threshold line (level of detection). In order to be plotted on the model surface (Fig. 3), averages of PCR results were expressed as 1.8^{Ct}, due to the rate of amplification efficiency of the PCRs (90%):

$$1{,}8^{\text{V-J quantif}}$$

Modeling approach: description of the stochastic model

Opening mechanism

The sequential windowing model we developed is a stochastic model consisting in two windows of accessibility progressing over the TRA/TRD locus over Vα and Jα regions. On the Figure 7, showing the diagram flow of the simulation program, the arrows represent the windows of accessibility moving over the V region from proximal V (3') to distal V genes (5') and over the J region from proximal J (5') to distal J genes (3'). For the first rearrangement round, the status of the V and J region accessibility window is calculated from the V and J opening speed parameters for both alleles. For the subsequent rearrangement rounds, the new sizes and positions of the V and J region accessibility windows result from the opening mechanism and from the DNA sequence(s) deleted by the previous rearrangement(s). The lengths of windows of accessible DNA over the V and J regions, LV_k and LJ_k verify the equations:

$$LV_1 = S_V\,(t_0 + \tau_1),\ LJ_1 = S_J\,(t_0 + \tau_1)$$

and, for k≥2:

$$LV_k = LV_{k-1} + S_V\,\tau_k,\ LJ_k = LJ_{k-1} + S_J\,\tau_k,$$

where k refers to the rearrangement round number. The opening offset time t_0 denotes a minimal time of opening and τ_k ($k\geq1$) are random variables uniformly distributed between t_0 and the end of the opening process. The V and J opening speed parameters, S_v and S_J, consist in varying parameters; their sensitivity was studied by simulations. The sequential windowing model was first developed in mouse. In this species, the V and J opening speed parameters were determined from the time of first detection of rearranged genes during ontogeny. They were described as non-constant speeds varying within an interval. In base of this experimental result, we used non-constant speeds as well for the opening speed parameters in the model developed for human TRA rearrangements. The quality of the fit of model results with experimental V-J profiles allowed evaluating the opening speeds values in human. The opening duration before each rearrangement is a constant parameter whose value was determined through simulations as well, by varying values between 2 hours and 50 hours.

Accessible genes and genes to rearrange

In the model, the position and size of the accessibility windows at the very time of a rearrangement event directly determine the sequential V and J genes being accessible for this rearrangement based on their physical positions. At each rearrangement event and for each allele, the model settles on a number of sequential V genes and J genes considered as accessible genes for the rearrangement. For each rearrangement round k, and for each such V_{ik} and J_{jk} genes, Boolean variables named BV_{ik} and BJ_{jk} were defined: they equal to 1 if the gene is accessible, and to 0 if it is non-accessible or deleted during a previous rearrangement of order less than k. Among the accessible genes, one V gene and one J gene are chosen according to the probability distribution given by their RSS scores. The RSS scores (KV_i and KJ_j) represent, for each V_i or J_j genes, the homology percentage of the gene sequence compared to the consensus sequence. The RSS scores take a value between 0 and 1, the 1 value corresponding to a fully consensus RSS. The RSS score we determined from homology percentages fell in agreement with the functional/pseudo gene status of all the V or J genes in human (Glusman *et al*, 2001; *Lefranc et al, 2009;* see www.imgt.org for RSS sequences). According to its non-functional status, a pseudo recombination gene is never found rearranged and consequently the corresponding RSS score is assimilated to zero in simulations. We denote by FV_k (resp. FJ_k) the cumulative distribution function of the relative window lengths obtained after the k^{th} rearrangement by adding the BV_{mk} (resp. BJ_{mk}) variables:

$$FV_k(i)=\Sigma_{m=1,\ldots,i} KV_m BV_{mk}/(\Sigma_{m=1,\ldots,104} KV_m\, BV_{mk})$$

$$FJ_k(j) = \Sigma_{m=1,\ldots,j} KJ_m BJ_{mk}/(\Sigma_{m=1,\ldots,60} KJ_m\, BJ_{mk})$$

At step k, we use the distribution functions FV_k and FJ_k corresponding to the random variables RV_k (resp. RJ_k) uniform on [0,1] and we calculate a number NV_k (resp.NJ_k) equal to:

$$\inf(FV_k^{-1}(RV_k)),$$
$$\text{and respectively } \inf(FJ_k^{-1}(RJ_k)),$$

where NV_k and NJ_k correspond to the V and J genes to rearrange at a step k.

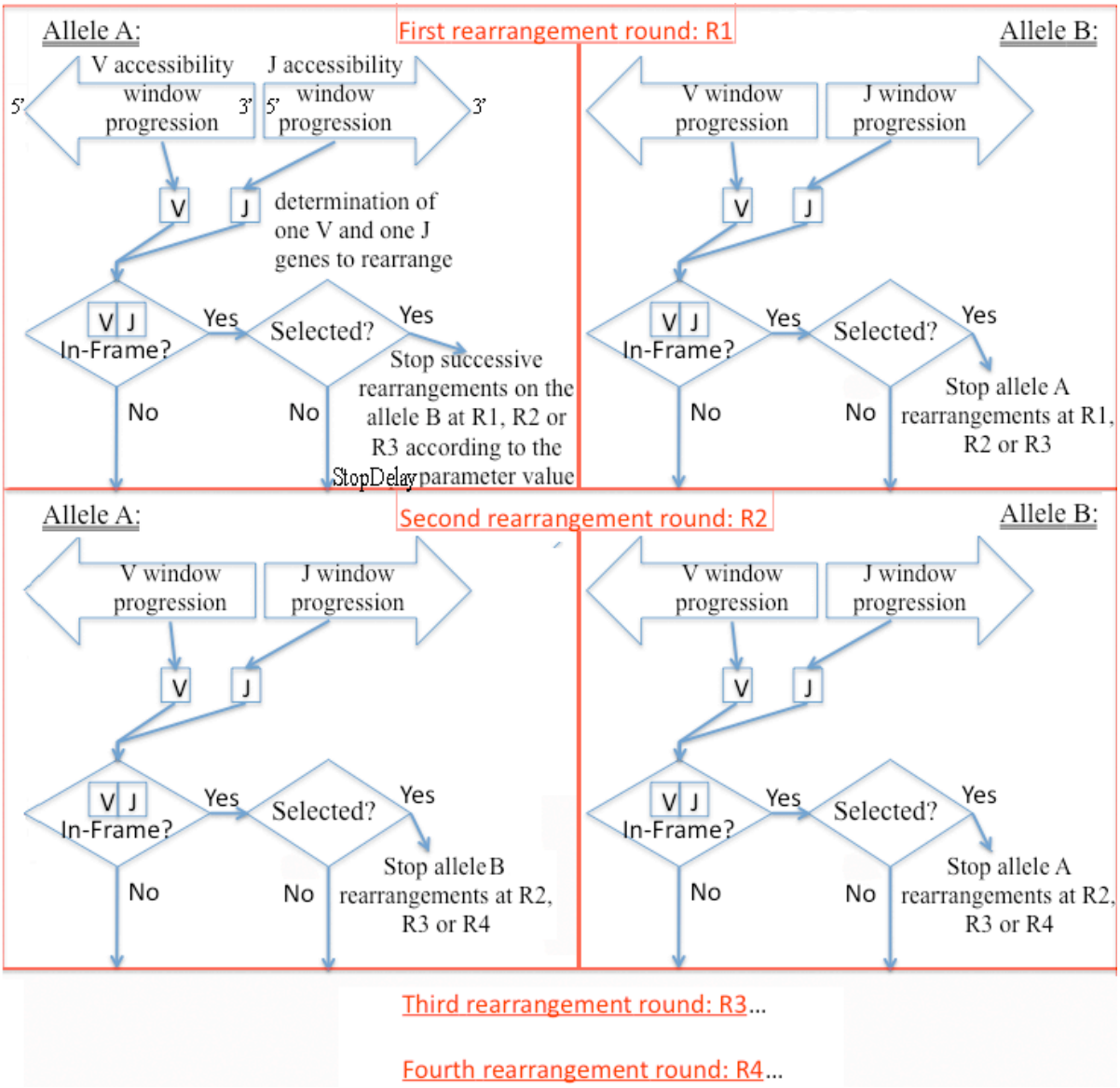


Figure 6: Stochastic successive windowing model diagram flow. This diagram flow represents, for each allele (named allele A and allele B), the succession of the rearrangement occurrences. Only the two first rounds of rearrangements are fully represented. The two arrows represent the progression of the accessibility windows over the V region (from proximal V genes at the 3' end to distal V genes at the 5' end) and over the J region (from proximal J genes at the 5' end to distal J genes at the 3' end). For each rearrangement occurrence, the sizes and positions of both windows are calculated by the program in function of parameters like the speeds of accessibility progression for the V and J region and according to DNA sequences that might be deleted from a previous rearrangement. Inside the windows, one V and one J are randomly chosen according to RSS score probabilities, which are determined from the percentage of homology of the RSS with the RSS consensus sequence (in the scheme: determination of one V and one J genes to rearrange). One V and one J genes to rearrange are determined, and the in-frame status in randomly fixed according to the IF parameter that can take a maximum value of 1/3. When an in-frame rearrangement is generated in the model, a global selection coefficient is randomly applied. Each selected in-frame rearrangement will stop the successive rearrangement mechanism in the whole cell according to the $Stop_{delay}$ parameter value (detailed in the model parameter section).

In-frame/out-of-frame rearrangements

The conservation of the proper reading frame of the genes is equiprobably determined. The in-frame rearrangement frequency parameter (if) gives the probability to perform a productive (in-frame) rearrangement. The frequency of in-frame rearrangements is known to be at maximum of 1/3 (Coleclough *et al*, 1983). This upper limit was used in simulations. Effects of lower values for the "if" parameter were tested by simulations as well.

Selection

When an in-frame rearrangement is generated on one or on both alleles, it/they undergo selection. The selection coefficient, s, included in our model refers to a global selection coefficient corresponding to the successive steps of selection: β pairing, surface expression efficiency and thymic selection. The sensibility of the "s" parameter was studied by simulations. We consider s as a constant parameter applied to every in-frame rearrangement. This homogenous re-shaping of the T cell repertoire gives statistically an evaluation of the frequency of each V-J association and was already successfully used in the Warmflash model (Warmflash *et al*, 2006). In fact, the distinction between CD4+ and CD8+ subsets is not supported in the experimental data.

Subsequent rearrangement rounds

If no in-frame rearrangement is selected on both alleles, another rearrangement round occurs in the program. The new sizes and positions of the V and J region windows are calculated as a result of the opening mechanism and of the DNA sequences deleted by the previous rearrangement. The successive rearrangement mechanism continues until an in-frame rearrangement is selected or until k, the index of rearrangement round R_k, reaches the number maximal of successive rearrangements, R_{max}. Different values of R_{max} were tested by simulations.

Stop of the successive rearrangement mechanism

Once an in-frame rearrangement is selected, it will block the successive rearrangement mechanism within the cell. A parameter named $Stop_{delay}$ is defined which takes integer values. If $Stop_{delay}=0$, the rearrangements will stop within the same rearrangement round as for the other allele.

For example, if an in-frame rearrangement is performed at R_1 on the allele A and successfully selected (Fig. 1), it will block rearrangements on the other allele B at R_1, and the data corresponding to the R_1 rearrangement round for allele B (i.e., V_i and J_j genes implicated, and the in-frame or out-of-frame status of the new gene built from this (V_i, J_j) rearrangement) will be stored in the simulation results. If $Stop_{delay}=1$, the rearrangements will stop within the subsequent rearrangement round for the other allele B, namely at R_2. More generally, we have:

$$R_{k\ alleleB}=R_{k\ alleleA}+ Stop_{delay}$$

If the first in-frame rearrangement being selected occurs over the B allele locus, we have:

$$R_{k\ alleleA}=R_{k\ alleleB}+ Stop_{delay}$$

The effect of the $Stop_{delay}$ parameter value on simulated results was studied using different parameter values and in a second time, using $Stop_{delay}=0$ and $Stop_{delay}=1$ inside different fractions of the entire simulated population. Experimental interallelic distance distributions in the Jα usage were employed to set the intervals of variation of the parameter values (Davodeau *et al*, 2001).

References

Aude-Garcia, C., M. Gallagher, P.N. Marche, and E. Jouvin-Marche. Preferential ADV-AJ association during recombination in the mouse T-cell receptor alpha/delta locus. Immunogenetics. 2001. 52:224-230.

Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu YK. PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res. 2009. 37:815-24.

Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. Science. 1999. 286: 958-961.

Azuma T, Igras V, Reilly EB, Eisen HN. Diversity at the variable-joining region boundary of lambda light chains has a pronounced effect on immunoglobulin ligand-binding activity. Proc. Natl. Acad. Sci. U. S. A. 1984. 81: 6139-6143.

Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, et al. IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Nucleic Acids Res. 2004. 32: D51–54.

Bassing CH, Swat W, Alt FW.The mechanism and regulation of chromosomal V(D)J recombination. Cell. 2002. 109: S45-55.

Bleakley K, Lefranc MP, Biau G. Recovering probabilities for nucleotide trimming processes for T cell receptor TRA and TRG V-J junctions analyzed with IMGT tools. BMC Bioinformatics. 2008. 9: 408.

Casanova JL, Romero P, Widmann C, Kourilsky P,. T cell receptor genes in a series of class I major histocompatibility complex-restricted cytotoxic T lymphocyte clones specific for a Plasmodium berghei nonapeptide: implications for T cell allelic exclusion and antigen-specific repertoire. J. Exp. Med. 1991. 174: 1371-1383.

Coleclough C. Chance, necessity and antibody gene dynamics. Nature. 1983. 303: 23-26.

Davodeau F, Difilippantonio M, Roldan E, Malissen M, Casanova JL, Couedel C, Morcet JF, Merkenschlager M, Nussenzweig A, Bonneville M, Malissen B. The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. EMBO J. 2001. 20: 4717-4729.

Fuschiotti P, Pasqual N, Hierle V, Borel E, London J, Marche PN, Jouvin-Marche E. Analysis of the TCR alpha-chain rearrangement profile in human T lymphocytes. Mol. Immunol. 2007. 44: 3380-3388.

Gahéry-Ségard H, Jouvin-Marche E, Six A, Malissen M, Cazenave P-A, and Marche P-N. Germ-line genomic structure of the B10. A mouse V alpha2 gene subfamily. Immunogen. 1996. 44:298-305.

Gilfillan S, Dierich A, Lemeur M, Benoist C, Mathis D. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. Science. 1993. 261: 1175-1178.

Hawwari A, Bock C, Krangel MS. Regulation of T cell receptor alpha gene assembly by a complex hierarchy of germline Jalpha promoters. Nat. Immunol. 2005. 6: 481-489.

Jouvin-Marche, E., C. Aude-Garcia, S. Candéias, E. Borel, M. Malissen, S. Hachemi-Rachedi, H. Gahéry-Ségard, P.-A. Cazenave, and P.N. Marche. 1998. Differential chronology of TCRADV2 gene use by alpha and delta chains of the mouse T cell receptor. Eur. J. Immunol. 1998. 28:818-827.

Krangel MS Mechanics of T cell receptor gene rearrangement. Curr. Opin. Immunol. 2009. 21: 133–139.

Krangel MS, Carabana J, Abbarategui I, Schlimgen R, Hawwari A. Enforcing order within a complex locus: current perspectives on the control of V(D)J recombination at the murine T-cell receptor alpha/delta locus. Immunol. Rev. 2004. 200: 224-232.

Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009. 37:1006-1012.

Jouvin-Marche E, Fuschiotti P, Marche PN. Dynamic aspects of TCRalpha gene recombination: qualitative and quantitative assessments of the TCRalpha chain repertoire in man and mouse. Adv. Exp. Med. Biol. 2009. 650: 82-92.

Malissen M, Trucy J, Jouvin-Marche E, Cazenave PA, Scollay R, Malissen B. Regulation of TCR alpha and beta gene allelic exclusion during T-cell development. Immunol. Today. 1992. 13: 315-322.

Oettinger MA, Schatz DG, Gorka C, Baltimore D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science. 1990. 248: 1517-1523.

Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche PN, Jouvin-Marche E. Quantitative and qualitative changes in V-J alpha rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor alpha chain repertoire. J. Exp. Med. 2002. 196: 1163-1173.

Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, et al. Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. J. Exp. Med. 1993. 178: 615–622.

Rast JP, Anderson MK, Strong SJ, Luer C, Litman RT, Litman GW. alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny. Immunity. 1997. 6: 1-11.

Takeshita, S., Toda, M. & Ymagishi, H. Excision products of the T cell receptor gene support a progressive rearrangement model of the alpha/delta locus. EMBO J. 1989. 8: 3261-3270.

Thompson SD, Pelkonen J, Rytkonen M, Samaridis J, Hurwitz JL Nonrandom. rearrangement of T cell receptor J alpha genes in bone marrow T cell differentiation cultures. J. Immunol. 1990. 144: 2829–2834.

Thuderoz F, Simonet MA, Hansen O, Pasqual N, Dariz A, Baum TP, Hierle V, Demongeot J, Marche PN, Jouvin-Marche E. Numerical modelling of the V-J combinations of the T cell receptor TRA/TRD locus. PLoS Comput. Biol. 2010. 6: e1000682.

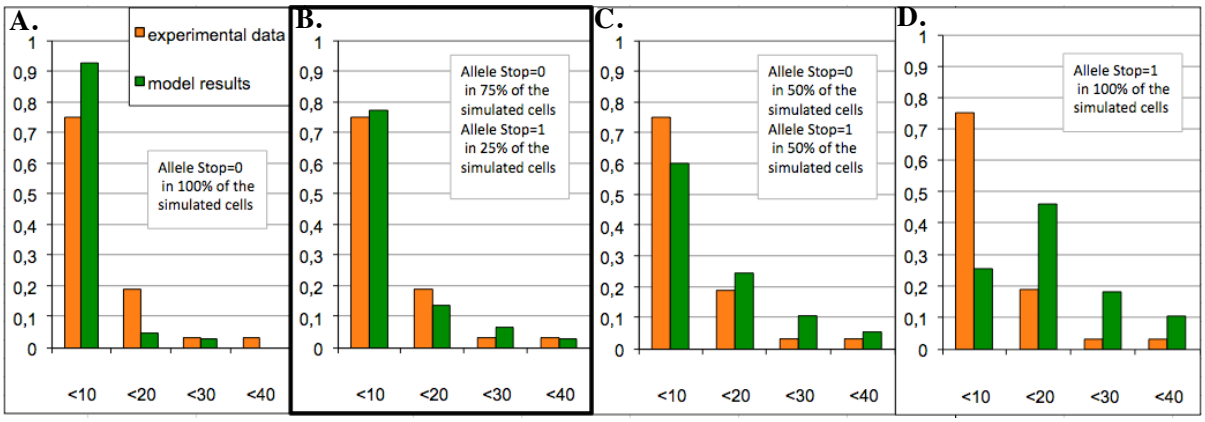
Uenishi H, Hiraiwa H, Yamamoto R, Yasue H, Takagaki Y, Shiina T, Kikkawa E, Inoko H, Awata T. Genomic structure around joining segments and constant regions of swine T-cell receptor alpha/delta (TRA/TRD) locus. Immunology. 2003. 109: 515-526.

Villey I, Caillol D, Selz F, Ferrier P, de Villartay JP. Defect in rearrangement of the most 5′ TCR-J alpha following targeted deletion of T early alpha (TEA): implications for TCR alpha locus accessibility. Immunity. 1996. 5: 331-342.

Warmflash A, Dinner AR. A model for TCR gene segment use. J. Immunol. 2006. 177: 3857-3864.

Wilson A, MacDonald HR, Radtke F. Notch 1-deficient common lymphoid precursors adopt a B cell fate in the thymus. J. Exp. Med. 2001. 194: 1003–1012.

Additional Figure



Additional Figure 1. Experimental J interallelic distance distribution in the mouse species presented along with mouse model results generated using different values for the Stop Delay parameter. The experimental distribution reproduced on the four graphs was compiled from four team studies addressing Jα usage at both TRA alleles, available in the literature, in order to form a more significant database (n=110), giving a statistically more confident empirical distribution of interallelic distances in J usage than in the human case (n=29) (Davodeau, 2001; Heath, 1995; Malissen, 1992; Casanova, 1991). The four graphs compare the experimental distribution with parametric study results generated varying the Stop Delay parameter value in mouse model. Comparison between experimental distribution and model simulation results showed a stop of successive rearrangements within the same rearrangement round ($Stop_{delay}$=0) and gave interallelic usage of some J genes too close compared to experimental distribution (A). On the contrary, a systematic stop of successive rearrangements within the subsequent rearrangement round ($Stop_{delay}$=1) gave interallelic usage of some J genes too distant compared to the experimental distribution (D). Finally, parametric studies along with a precise experimental distribution (calculated from a sample of size n=110), allow to predict that when an in-frame rearrangement is successfully selected in a cell, the intracellular signaling will stop further rearrangements over the other allele during the same rearrangement round for roughly 75% of the cells and during the subsequent rearrangement round for the rest of the cells (B, rejecting the C configuration).

Réseaux génétiques du système immunitaire

Ce dernier chapitre a pour but de tracer une brève synthèse entre le savoir mathématique développé à propos des réseaux d'interaction à un niveau théorique par d'autre doctorants de notre laboratoire et le savoir actuellement disponible dans la littérature sur les réseaux génétiques impliqués dans la différenciation des cellules du système immunitaire. La spécificité des réseaux génétiques contrôlant les réarrangements des loci de chaine de récepteurs antigéniques est la nécessité de mener une remodélisation de la chromatine et une relocalisation du locus avant les réarrangements. Cette caractéristique est nouvelle par rapport à d'autres réseaux de gène car elle correspondrait à des modes d'implémentation drastiquement différents à l'intérieur d'un même réseau. Conséquemment ceci fait émerger des axes de recherche intéressants dans leurs aspects mathématiques théoriques à propos des comportements de réseaux et pour l'interprétation des réseaux immuns également. Sous ce point de vue, le présent travail de thèse forme partie intégrante d'intérêts de recherche plus globaux de l'équipe Génome du laboratoire TIMC-IMAG. De la même manière, à la fin de ce chapitre, un parallèle est établit entre les réseaux immuns et le vieillissement qui constitue le sujet de recherche du laboratoire AGIM nouvellement formé.

This last chapter aims to draw a brief synthesis between the mathematical knowledge about regulatory network reached at a theoretical level by other PhD students of our laboratory and currently available knowledge about immunological networks from literature. The specificity of the immune network that controls rearrangements of antigenic receptor loci consists in the necessity to perform chromatin remodeling and locus relocalisation previously to rearrangements. This represents a new feature compared to other genetic networks because it would correspond to drastically differential incrementation modes within a same network, and therefore points out interesting research axes for the mathematical theoretical aspects about network behaviors and for the immune network interpretations as well. In this way, the present PhD work forms part of more global research interests of the TIMC-IMAG laboratory Human Genomics team. In the same way, at the very end of the chapter a parallel between immune network and ageing, which constitutes the investigation topic of the newly formed AGIM laboratory, is outlined.

Biological regulatory networks

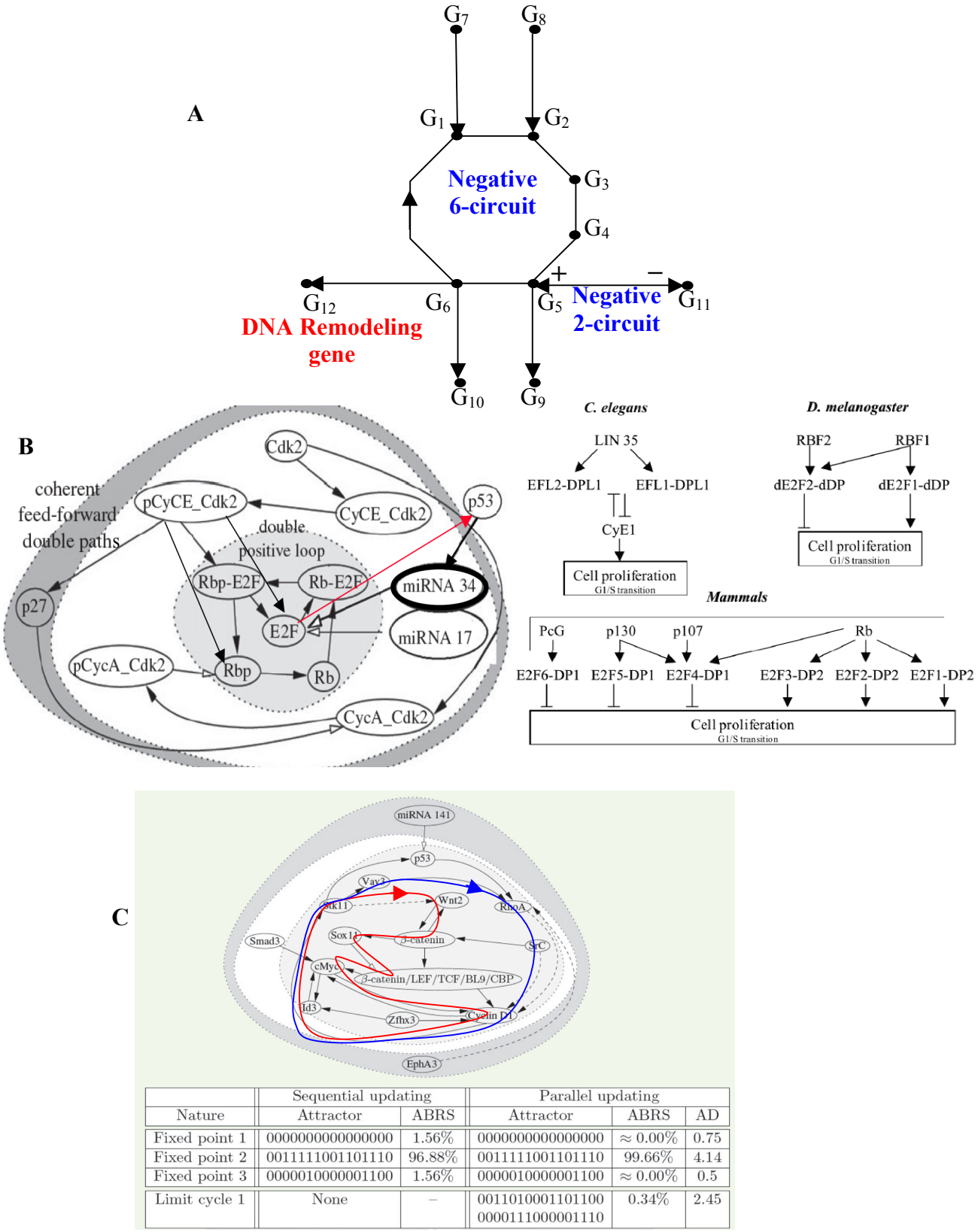
Generalities

The theory of biological regulatory networks was born in parallel to those about neural networks [Delbrück *et al*, 1949; Thomas *et al*, 1980; Kauffman *et al*, 1993] and currently represents a research topic of intense activity for interpreting the "omic" data from bio-arrays devices. The networks are made of elements in interaction: genes, proteins, or neurons. The central tool used for representing these interactions consists in a directed graph, called interaction graph, whose signed arrows (positive or negative) are related to the activation /inhibition influence exerted by inducer/repressor elements on others. For example, the introduction of inhibitions in a biological regulatory network can be made through miRNAs; in this case, inhibitions come from the sources of the interaction graph, also called upper nodes of the up-trees, and converge toward the "core" of the network (in the *"graph sense"* according to [Berge *et al*, 1974; Bollobas *et al*, 1985]), as illustrated by the G_7 and G_8 nodes showed on the Figure 1.A. Endogeneous repressors coming from the core of the graph can also exert inhibitions. The networks core is compound of intersecting circuits, a circuit of size n being constituted by an interaction loop between consecutive genes, starting at a gene G_1, acting on a gene G_2, acting on G_3, … , and eventually acting on the terminal gene of the loop G_n, which acts itself on G_1, the very first gene of the loop. Figure 1.A illustrates a regulatory network core made of two intersecting circuits, both being supposed negative, for they include an odd number of inhibitions (circuits presenting an even number of negative interactions being positive).

The notion of attractor

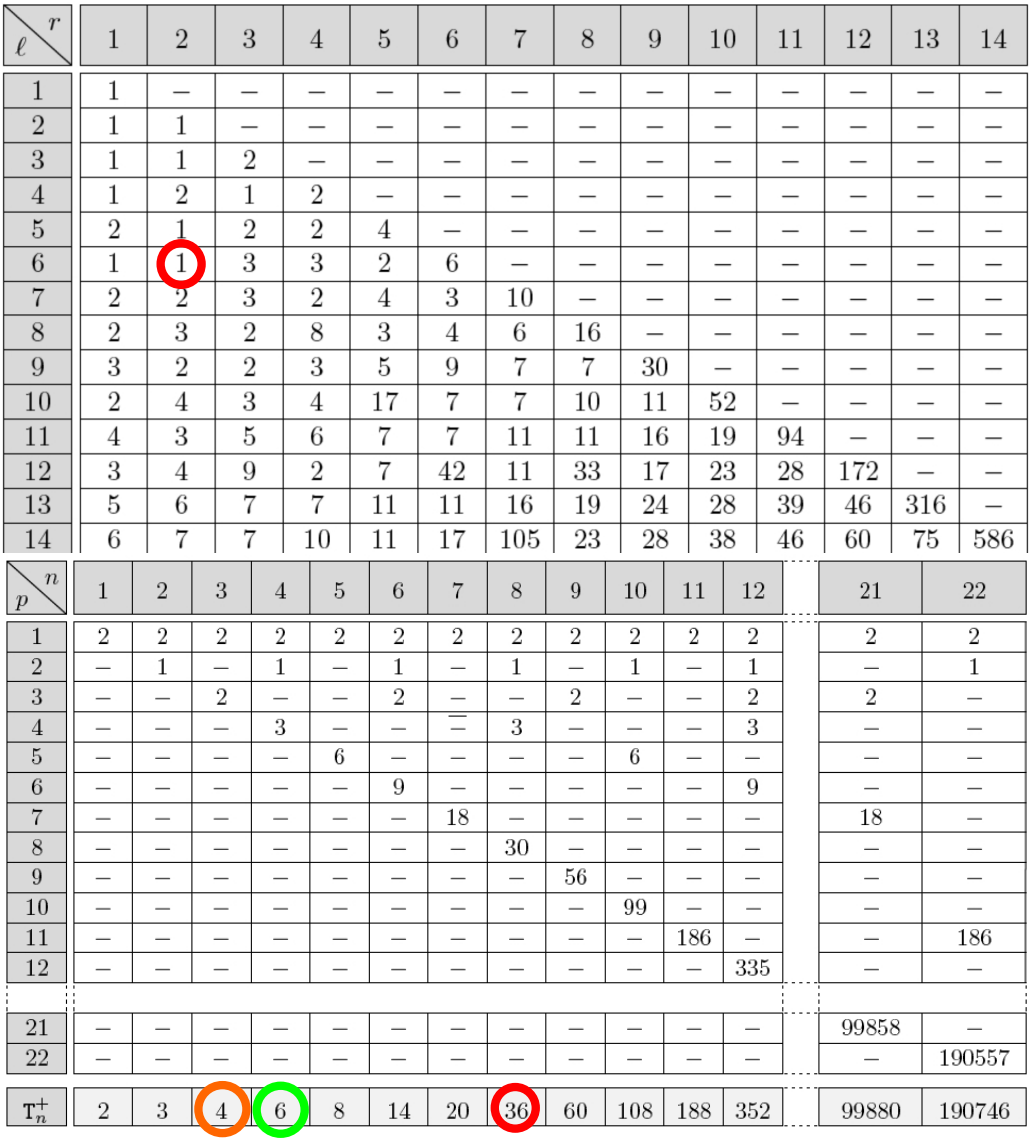
For each given network, the number of asymptotic steady behaviors (in time) can be calculated; such states are named attractors and identified as the possible differentiated cell states of a tissue [Delbrück *et al*, 1949]. A huge effort was accomplished in mathematical research in order to identify the source of the attractor multiplicity (which turned out to be closely related to the number of positive circuits in the network) [Demongeot *et al*, 1988; Snoussi *et al*, 1998; Cinquin *et al*, 2002], or on the contrary, to identify the circuit configurations related to the attractor uniqueness. These two aspects present a high interest for ensuring a sufficient number of differentiated functions (about 300 in human [Demongeot *et al*, 2009]) or for providing a tissue devoted to a unique function.

It is worth noticing that if the G_5 gene in Figure 1.A, which is common to two negative circuits, is decoupled into two gene interactions, the circuit becomes positive. Indeed, the circuit would be in this case composed by an even number of negative interactions, corresponding to the sum of the two odd numbers of inhibitions of each previous negative circuit. This positive circuit of length 8 in our example displays 36 attractors in parallel mode of updating, of which 30 are limit cycles of order 8 (cf. Figure 2 bottom, red circle, from theoretical results). Thus, the functional role of the intersection of negative circuits would consists in drastically reducing the number of possible attractors of the network, making it completely stable (in our example) and devoted to only one function. In fact, from the Figure 2 top (red circle), it can be determined that the intersection of 2 negative circuits, one of length 6 and the other of length 2, presents only one attractor. The attractor number values predicted from mathematical calculations were gathered and presented in Figures 2; these theoretical results are in accordance with numerical simulation results.



	Sequential updating		Parallel updating		
Nature	Attractor	ABRS	Attractor	ABRS	AD
Fixed point 1	0000000000000000	1.56%	0000000000000000	≈ 0.00%	0.75
Fixed point 2	0011111001101110	96.88%	0011111001101110	99.66%	4.14
Fixed point 3	0000010000001100	1.56%	0000010000001100	≈ 0.00%	0.5
Limit cycle 1	None	–	0011010001101100 0000111000001110	0.34%	2.45

Figure 1. Regulatory networks. A. Theoretical network with two negative circuits of respective 6 and 2 lengths, sharing the node G_5. **B.** Graph core of the regulatory network controling the cell cycle in mammals (left) and evolution of the up-tree over two positive circuits sharing two nodes. After [Demongeot *et al*, 2009]. **C.** Top: Regulatory network of the feather morphogenesis with a positive circuit of length 4 (blue) and a negative of length 7 (red). Bottom: explicitation of the attractors of the network dynamics. After [Demongeot *et al*, 2009].



ℓ \ r	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	–	–	–	–	–	–	–	–	–	–	–	–	–
2	1	1	–	–	–	–	–	–	–	–	–	–	–	–
3	1	1	2	–	–	–	–	–	–	–	–	–	–	–
4	1	2	1	2	–	–	–	–	–	–	–	–	–	–
5	2	1	2	2	4	–	–	–	–	–	–	–	–	–
6	1	1	3	3	2	6	–	–	–	–	–	–	–	–
7	2	2	3	2	4	3	10	–	–	–	–	–	–	–
8	2	3	2	8	3	4	6	16	–	–	–	–	–	–
9	3	2	2	3	5	9	7	7	30	–	–	–	–	–
10	2	4	3	4	17	7	7	10	11	52	–	–	–	–
11	4	3	5	6	7	7	11	11	16	19	94	–	–	–
12	3	4	9	2	7	42	11	33	17	23	28	172	–	–
13	5	6	7	7	11	11	16	19	24	28	39	46	316	–
14	6	7	7	10	11	17	105	23	28	38	46	60	75	586

p \ n	1	2	3	4	5	6	7	8	9	10	11	12	21	22
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	–	1	–	1	–	1	–	1	–	1	–	1	–	1
3	–	–	2	–	–	2	–	–	2	–	–	2	2	–
4	–	–	–	3	–	–	–	3	–	–	–	3	–	–
5	–	–	–	–	6	–	–	–	–	6	–	–	–	–
6	–	–	–	–	–	9	–	–	–	–	–	9	–	–
7	–	–	–	–	–	–	18	–	–	–	–	–	18	–
8	–	–	–	–	–	–	–	30	–	–	–	–	–	–
9	–	–	–	–	–	–	–	–	56	–	–	–	–	–
10	–	–	–	–	–	–	–	–	–	99	–	–	–	–
11	–	–	–	–	–	–	–	–	–	–	186	–	–	186
12	–	–	–	–	–	–	–	–	–	–	–	335	–	–
21	–	–	–	–	–	–	–	–	–	–	–	–	99858	–
22	–	–	–	–	–	–	–	–	–	–	–	–	–	190557
T_n^+	2	3	4	6	8	14	20	36	60	108	188	352	99880	190746

Figure 2. Calculations of the number of attractors in intersecting negative circuits of size *r* and *l* (top) and of positive circuits of size *n*, having 1, 2, ... , p elements, the last row giving the total number attractors (bottom).

In order to provide more illustrations of this theoretical statement, Figure 1.B presents a genetic network controlling cell proliferation. According to the theoretical results mentioned, this network, including intersecting circuits, displays only 1 attractor in parallel updating mode for the intersecting positive (size 3) and negative (size 3) circuits but gives 4 attractors if the positive circuit is alone (Figure 2 bottom and Figure 13 top, orange circles). On another example, the Figure 1.C presents the network controlling the feather morphogenesis. This network includes a negative circuit of length 7 intersecting on 2 nodes a positive circuit of length 4, and thus would display less than 3 attractors (because 3 is the number of attractors when the circuits intersect on

one node only, cf. Figure 13 middle, green circle), but the same network would exhibit 6 attractors if the positive circuit is alone (Figure 2 or 13 bottom, green circle). On a last example, in the operon regulating the *Arabidopsis* thaliana flower morphogenesis [Mendoza *et al*, 1998], the interaction matrix is a (11,11)-matrix with only 22 non-zero coefficients (see Figure 6 of [Demongeot *et al*, 2002], which is presented in the Annex 1 of the present chapter). The corresponding interaction graph presents P(W)=4 positive intersecting circuits as well as A(W)=4 attractors (in parallel mode of updating) displaying a "sufficiently" large attraction basin to be observable, corresponding to the exact number needed to ensure flowering functions [Mendoza *et al*, 1998].

The considerations made above about genetic networks illustrate how the concept of attractor constitutes the most important notion when addressing biological regulatory networks. The network attractors are studied through an interaction matrix W, which is associated to the interaction graph (the "incidence" matrix). The interaction matrix is similar to the synaptic weight matrix in neural networks, which stands for the relationships between neurons. The general coefficient w_{ik} of the W interaction matrix is equal to +1 (and respectively –1 and 0) if the G_k gene activates (resp. inhibits and does not influence) the G_i gene. The x_i state of the G_i gene is equal to +1 (resp. –1), if the gene is (resp. is not) expressed. The state change of a given G_i gene between t and t+1 obeys a threshold rule:

$$x_i(t+1)=H(\sum_{k=1,n} m_{ik}x_k(t)-b_i) \text{ or } x(t+1)=H(Wx(t)-b),$$

where H is the sign step function (H(y)=1, if y≥0 and H(y)=-1, if y<0), and b_i's are the threshold values. When t increases, the states of the genes reach a stable set of configurations (a fixed configuration or a cycle of configurations), which constitutes an attractor of the genetic network dynamics. The attractor attraction basin corresponds to a set of states leading inevitably to the attractor state(s) after iteration of the interaction rules. By changing the sign of the state variable x_i into the associated Boolean variable $y_i = (x_i+1)/2$, we obtain the so-called Boolean regulatory networks, which have the same attractors than the corresponding signed networks.

In the case of organisms, whose the simple genome corresponds to small regulatory genetic systems, knowledge about the W matrix permits to explicit the entire stationary asymptotic behavior possibilities. The calculation of the number of attractors of more complex biological dynamical systems constitutes a difficult problem addressed by mathematical approaches [Demongeot *et al*, 1988; Snoussi *et al*, 1998; Cinquin *et al*, 2002; Demongeot *et al*, 2009], and remaining open in most of its applications. In general, it results of great biological interest and relevance to determine interaction matrices possessing characteristic properties as i)

a minimal number of non zero coefficients for a given set of attractors or ii) a minimal number P(W) of positive loops, controlling the number A(W) of attractors and their stability (cf. [Demongeot *et al*, 1988; Snoussi *et al*, 1998; Demongeot *et al*, 2009] for the continuous case and [Cinquin *et al*, 2002] for the discrete one). In the following, it is shown how former mathematical results partly solve these problems of minimality, by giving necessary and sufficient conditions to obtain the i) and ii) properties. In order to calculate the w_{ik}'s, we can first estimate the s-directional correlation $\rho_{ik}(s)$ between the state vector $\{x_k(t-s)\}_{t\in C}$ of gene j at time t-s and the state vector $\{x_i(t)\}_{t\in C}$ of gene i at time t, t (bio-arrays observation times), varying during the time set C of the cell cycle:

$$\rho_{ik}(s)=(\sum_{t\in C} x_k(t-s)x_i(t)-[\sum_{t\in C}x_k(t-s)\sum_{t\in C} x_i(t)]/|C|)/\sigma_k(s)\sigma_i(0),$$

$$\text{where } \sigma_k(s)=(\sum_{t\in C} (x_k(t-s)^2-[\sum_{t\in C} x_k(t-s))^2]/|C|)^{1/2},$$

and then take $w_{ik}=\text{sign}(\sum_{s\in C} \rho_{ik}(s)/|C|)$, if $|w_{ik}|>\eta$, and $w_{ik}=0$, if $|w_{ik}|<\eta$, where η is a de-correlation threshold. We can also identify the system with a Boolean neural network. When the obtention of the entire coefficients of W is unfeasible (neither from the literature nor from such calculations), the missing coefficients would be chosen randomly by respecting the connectivity value K(W)=I/N (ratio between the number I of interactions and the number N of genes), and the mean inhibition weight I(W)=R/I (ratio between the number of inhibitions R and I). In fact, the K(W) value is in general comprised between 1.5 and 3, and I(W) normally stands between 1/3 and 2/3, as observed for many networks (lactose operon, Cro operon for the phage λ, lysogenic/lytic operon for the phage μ, gastrulation network, *Arabidopsis* thaliana flowering network,…). Calling G the interaction graph associated to W, the strongly connected components of G denotes the sets of genes such that for any pair of them, there is at least one non-interrupted connected path among the sequences of arrows of G connecting the genes of this pair. A source represents a gene receiving no arc, but influencing at least one other gene. A regulon denotes a minimal connected component of G having exactly 1 positive loop (auto-catalysis of one node) and 1 negative circuit, including the auto-catalysed node. In the lactose operon example, K(W)=8/6, I(W)=3/8, P(W)=2, A(W)=2 (βgal-activated and inactivated states), and G includes 1 connected component as well as 1 regulon.

Complementary definitions and notations

In the following, some definitions are presented about the rigorous mathematical description of the discrete Boolean networks used to describe the genes interaction dynamics, their associated interaction graph G, and the incidence matrix W. For this purpose, some theoretical results with rigorous proofs from [Demongeot *et* al, 2002; Demongeot *et* al, 2003] are introduced only for the first and last results in order to illustrate the kind of mathematical reasoning is performed. The complete demonstrations are available in the referred articles presented in Annex 2 and 3 of the present chapter.

Considering a digraph G=(V,E), where V={1,..., n} corresponds to the set of nodes and E states for the set of oriented arcs, let W=(w_{ik}) be a real (n,n)-matrix. W is the incidence matrix of G, if w_{ik} = +1 (resp. -1) 0, and if and only if (k,i) is a positive (resp. negative) arc going from k to i in E. By extension, G is also called the incidence graph of W. By definition of W, the sign of the arc (k,i), denoted by sign((k,i)) is also the sign of w_{ik}. Let us denote by $G^-(i)$ (resp. $G^+(i)$) the set of nodes $\{i_1, i_2,\ldots, i_{k(i)}\}$ such that (i_j,i) (resp. (i,i_j)) belongs to E, for each j=1,…,k(i). We will consider that a set of arcs C = $\{e_1, e_2,\ldots, e_r\}$ is a chain if each e_k in C has a node belonging to e_{k-1} and the other one belonging to e_{k+1}. We will consider that C is a simple (resp. elementary) chain if the arcs (resp. nodes) are different. In the sequel, we will understand by chain a simple and elementary chain. In the same way, we will call C a path if $e_k = (i_k,i_{k+1})$ implies $e_{k+1} = (i_{k+1},i_{k+2})$, for all k=1,…, r, that is to say the final node of each arc is the beginning node of the next arc in C. The sign of a path C (denoted by sign(C)) is positive if the number of negative arcs of C is even and negative if this number is odd. A circuit (or loop) is defined as a path whose any arc two extremities belongs to two and only two arcs. For simplicity of notation, we will say that a node i belongs to a circuit C if there exists a node j such that (i,j) or (j,i) belongs to C. Every other definition of graph theory used here will be consistent with that presented in [Berge *et al*, 1974; Bollobas *et al*, 1985]. A circuit C is called negative (resp. positive) if the sign(C) is negative (resp. positive). Let define now a discrete state regulatory network, acting on the set of states {-1,1}, as the 4-uple N=(G,W,b,sign), where G is the incidence graph of W, b is a threshold real vector and the local transition function is given by x_i (t+1)= sign ($\Sigma_{k=1,\ldots,n} w_{ik}x_k(t)-b_i$), $\forall$ x(t) $\in \{-1,1\}^n$, where sign(u) =1, if u $\geq$ 0, and sign(u) = -1 otherwise. Sequential iteration consists to update nodes one by one in a prescribed periodic order I=(i(1), i(2), …, i(n)), where {i(1), i(2), …, i(n)}={1, 2, …, n }, i.e., starting with x(0)=($x_1(0)$, …, $x_n(0)$) in $\{-1,1\}^n$, to generate the sequence of iterates:

$$x_{i(k)}(t) = \text{sign}\,(\Sigma_{j<k}\, w_{i(k)i(j)}x_{i(j)}(t) - b_{i(k)} + \Sigma_{j\geq k}\, w_{i(k)i(j)}x_{i(j)}(t-1) - b_{i(k)}),\ \forall\, k \in \{1, \ldots, n\}.$$

Now, the parallel iteration consists in updating all the nodes synchronously:

$x_i(t+1) = \text{sign}\,(\Sigma_{k=1,\ldots,n}\, w_{ik}x_k(t) - b_i)$, $\forall\, i \in \{1, \ldots, n\}$, with $x(0) \in \{-1,1\}^n$. We shall say that x is a fixed point if x is invariant under the application of the complete sequence of updates. It is important to observe that the kind of iteration does not change the set of fixed points, but only change their attraction basins. In the following we will use systematically the parallel iteration.

Relations between positive/negative circuits, and fixed points

In the sequel, we will assume that the graph G is strongly connected, and that therefore one can apply the results to each strongly connected component of G. In addition, we will suppose with not loss of generality that $|G^-(i)| > 0$, for all $i \in V$. If there exists a node $i \in V$ such that $G^-(i)$ is empty, then we can assume that the arc (i,i) exists in E; in this way, the dynamics of both networks are the same. It evidently follows from this property that there exists at least one circuit C in G (eventually a circuit of the form (i,i)). Eventually, we suppose that the graph G and the matrix W have a quasi-minimal structure, that is to say, all (j,i), such as $i \neq j$, belongs to E (or equivalently $w_{ij} \neq 0$, if $i \neq j$), if and only if there exists $x \in \{-1,1\}^n$, such that :

$$\text{sign}\Big(\sum_k w_{ik}x_k - b_i\Big) \neq \text{sign}\Big(\sum_{k\neq j} w_{ik}x_k - b_i\Big)$$

Hence, we have the following necessary condition to get a quasi-minimal structure for G:

$$-\sum_k |w_{ik}| < b_i \leq \sum_k |w_{ik}|, \quad \forall\, i=1,\ldots, n.$$

The next property will be useful in the following for characterizing a positive circuit:

Proposition 1: A circuit C is positive if and only if there exists a vector $x \in \{-1,1\}^n$ such that for all $(k,i) \in C$, $\text{sign}(w_{ik}) = x_i x_k$ or equivalently, for all $(k,i) \in C$, $x_i = \text{sign}(w_{ik})x_k$ (1)

Proof: Let C be a positive circuit and i(0) a fixed node belonging to C. Let us enumerate the nodes belonging to C by i(0), i(1),… , i(j), such that $\forall\, k = 0,\ldots, j$,
either (i(k-1),i(k)) or (i(k),i(k-1)) $\in$ C (by identifying j and –1). Finally, let us define the vector x as follows:

- $x_{i(0)} = 1$ and $x_{i(k)} = \text{sign}(w_{i(k)i(k-1)})x_{i(k-1)}$, if $(i(k-1),i(k)) \in C$ or
  $x_{i(k)} = \text{sign}(w_{i(k-1)i(k)})x_{i(k-1)}$, if $(i(k),i(k-1)) \in C$, $\forall\, k =1,\ldots, j$.

Obviously x is satisfying the equation (1); hence -x satisfies the equation (1) too. Finally, it is direct that there does not exist another vector $y \notin \{x, -x\}$ that satisfies the equation (1).

Let C be now a negative circuit, and let us suppose that the equation (1) is true, then $\prod_{(j,i)\in C} sign(w_{ij})$ $= \prod_j x_j \prod_i x_i = (\prod_j x_j)^2$, but $\text{sign}(C) = \prod_{(j,i)\in C} sign(w_{ij}) < 0$, which is contradictory.

We can therefore write the following theorems:

Theorem 1: Given N, if all circuits of the incidence graph G are positive, then, there exists a vector $x = (x_1 \ldots, x_n) \in \{-1,1\}^n$ such that x and $-x = (-x_1,\ldots,-x_n)$ are fixed points of N.

Remark: there is 2 remarkable fixed points having by construction a non frustration property, that is on each circuit, the sign changes of x_i's are identical to the sign changes of the arcs. For other eventual fixed points, there is at least 1 cycle for which sign changes are frustrated.

Theorem 2: if all circuits of the incidence graph G are negative, then N has no fixed points.

Minimal regulatory networks

The previous results allow us to characterize some minimal regulatory networks, and the following propositions will constitute minimal regulatory network examples. These propositions solve in part the inverse problem consisting in the description of W only from the knowledge of a phenotypic x observed from bio-arrays.

Proposition 2: let N having n nodes and n connections, a necessary and sufficient condition of existence of a fixed point x is the existence of a positive circuit. In this case, x and -x are both fixed points. Hence we can characterize the set of minimal N's having x as fixed point.

Proposition 3: given a state vector x, the set of minimal networks N=(G,W,b,sign) having x as fixed point is given by the following conditions:

$w_{ik} = a_{ik}x_i x_k$, where $a_{ik} \geq 0$ and, for all i, there exists k(i) such that $a_{ik(i)} \neq 0$

$-|a_{ik(i)}| < b_i \leq |a_{ik(i)}|$.

Conjecture: let N with n nodes (n>7) and n+1 connections, a necessary and sufficient condition for the existence of a unique attractor in parallel iterations, is a negative and a positive circuit intersecting, the length of the positive circuit dividing the length of the negative one. The unique attractor is a fixed point. Let us notice that the Conjecture has been partially proved in [Demongeot *et al*, 2003] (Annex 3), and is verified by the simulations of Figure 2 top and 13 top.

Fixed points bounds in regulatory networks

Given N and C a positive circuit of N, by Proposition 1, there is x in $\{-1,1\}^{|V(C)|}$, such that x and -x satisfy the equation: $\forall (k,i) \in C$, $sign(w_{ik})=x_i x_k$. By denoting $u(C) \in \{-1,1\}^{|V(C)|}$ the vector equal to x (resp –x), if $x_{i(0)}=1$ (resp –1), where $i(0)=\min\{i \mid i \in C\}$, we have:

Lemma 1: Given N and y a fixed vector of N, then for all $i \in V$, there exists a positive circuit C(i) in G such that for all k in C(i), $y_k=u(C(i))_k$ or for all k in C(i), $y_k=-u(C(i))_k$.

Theorem 3: If m is the total number of positive circuits of N, then the number of fixed points of N is $\leq 2^m$, and this upper bound is reached if and only if for all circuits C of N, there does not exist an arc (k,i) starting in C^C and ending in C (there is no source for C not located in C).

The "immunetworks"

The term "immunetworks" designates the biological regulatory networks involved in the control of the immune response. These networks are currently extensively studied and numerous genes and interactions composing their interaction graphs were identified.

Concerning the rearrangements of the TCRB and TCRA loci, this process is controlled by hundred of genes forming the immunetwork presented on the Figure 3, from [Georgescu *et al*, 2008]. The rapid analysis of the connected components of this interaction graph reveals structural similarities with the conceptual Figure 1.A network. In fact, this complex immunetwork possesses two intersecting negative circuits, one of length 6 (the blue circuit on Figure 3) and the other of length 2 (the green Ikaros circuit on Figure 3), intersected at the PU.1 gene node. Therefore, in the Figure 4, the conceptual genes of the Figure 1 were identified and instantiated following the structure of interaction of the immunetwork compiled by Georgescu *et al* [Georgescu *et al*, 2008]. As a first conclusion, the strongly connected component structure tells the existence of a unique attractor coming from the two intersecting circuits (cf. Figure 13 top) giving the value 1 for the RAG gene, that would allow performing first the TCRB rearrangements, and afterward the TCRA recombinations.

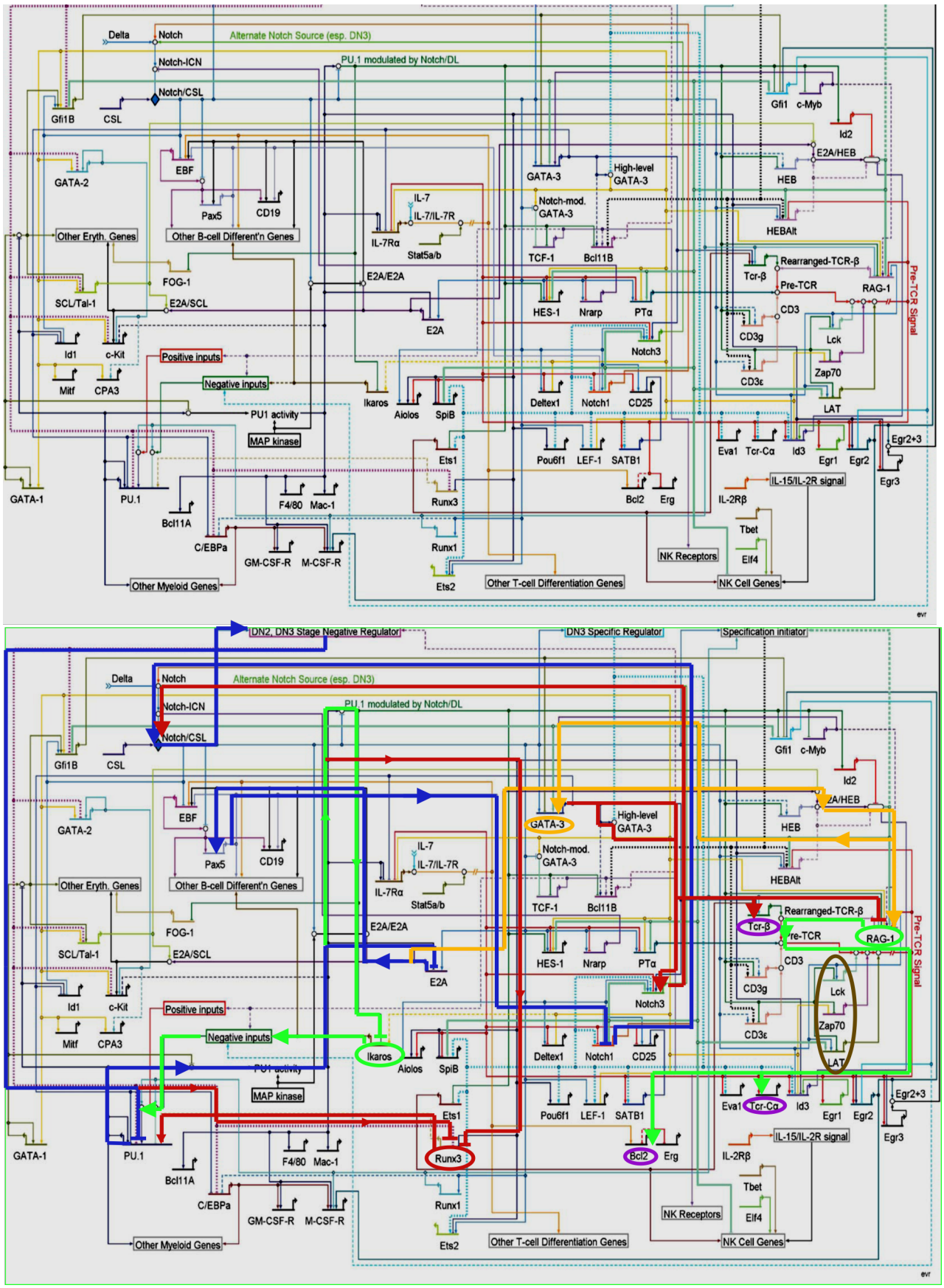


Figure 3. Regulatory network controlling the immune system. After [Georgescu *et al*, 2008].

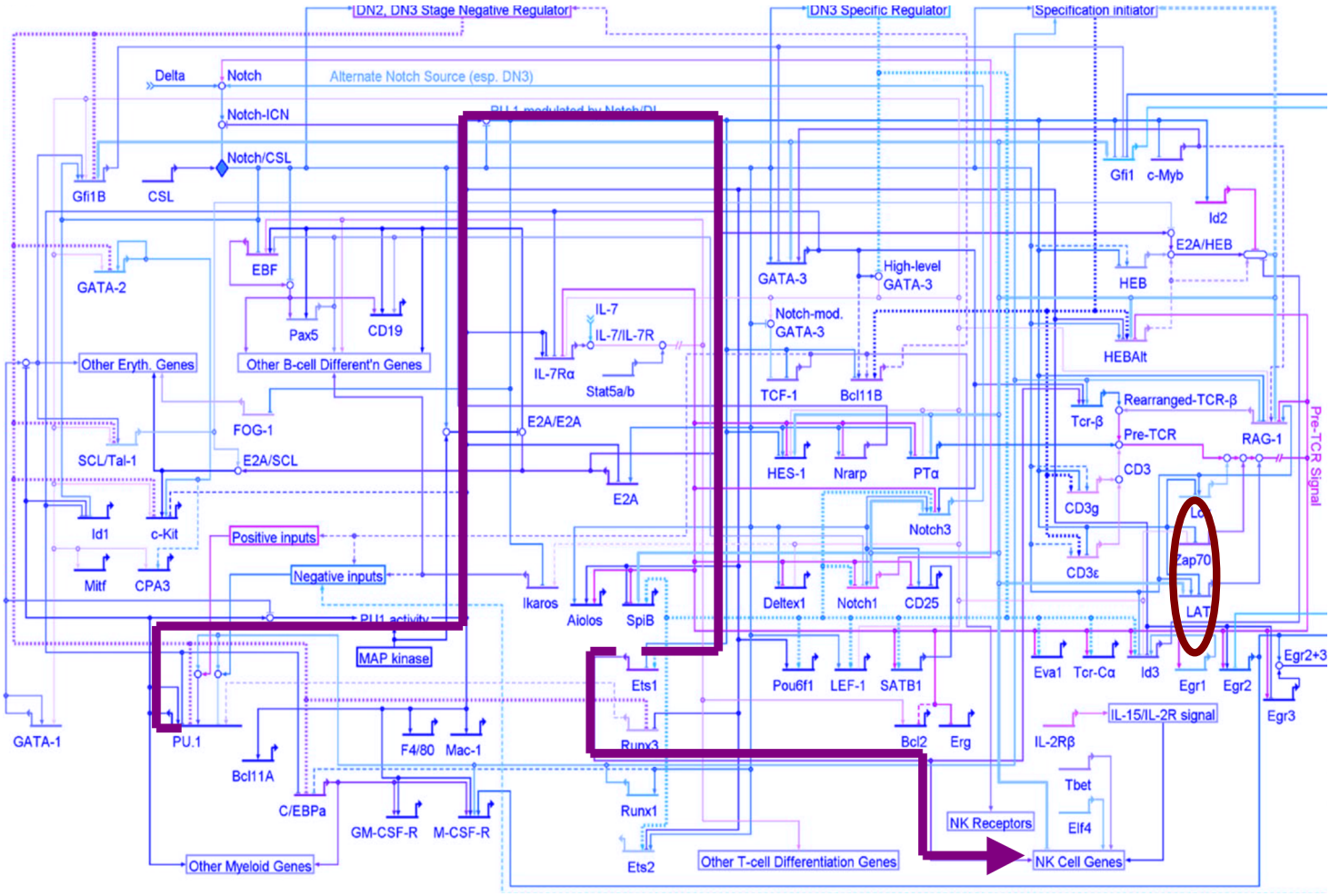


Figure 3 bis. The regulation of the NK cell genes by PU.1. After [Georgescu *et al*, 2008].

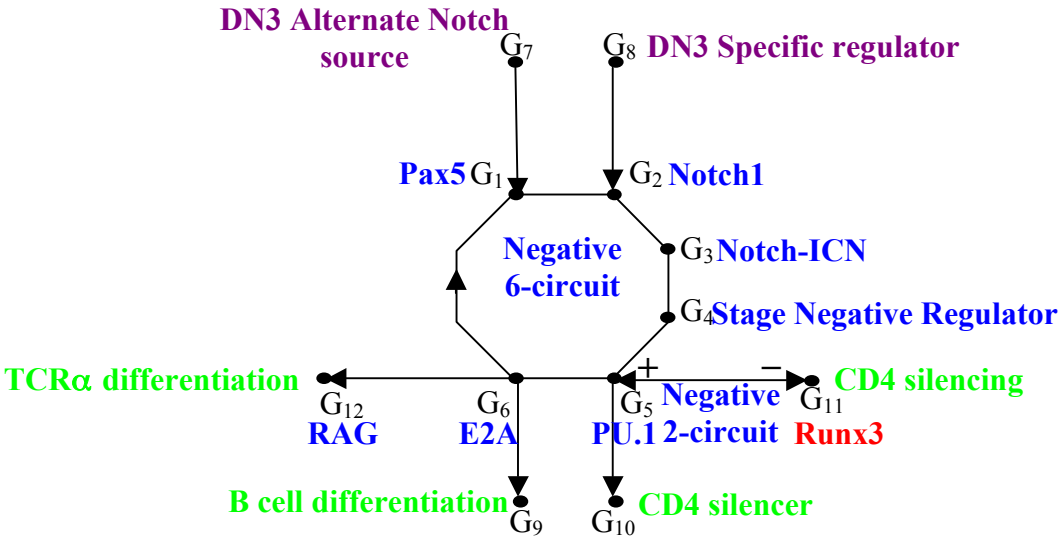


Figure 4. Instantiation of the genes of the Figure 1 from the regulatory network of the Figure 3

Interestingly, the immunetwork controlling the B cells differentiation from [Singh *et al*, 2005] and presented on Figure 5 involves strongly confirmed interactions (in full line), which are entirely shared with the immunetwork involved in the TCRA rearrangements (in blue on Figure 5, from Figure 3), showing a general use of successful solutions of regulation.

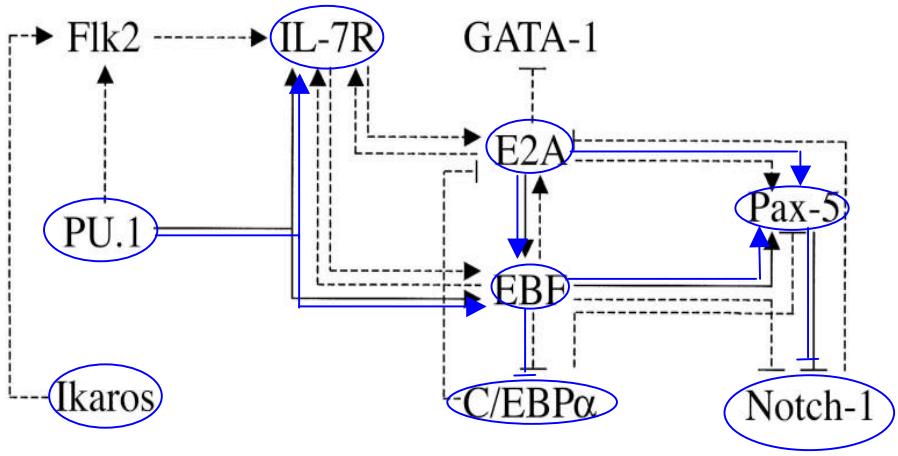


Figure 5. Genetic control common between the TCRA rearrangements (blue) and the B cells differentiation (black). After [Singh *et al*, 2005].

miRNAs implications in immunetworks

Typically, the most important interaction motifs ensuring network stability are the negative feedbacks, based on the presence of a negative circuit. Negative feedback requests the presence of an odd number of inhibitions in the circuit, hence the existence of an inhibitory activity. Inhibitory activities can also be due to microRNAS (miRNAs), which are ubiquitously present in all species, even in pathogens like viruses, as recently discovered in herpes viruses [Stern-Ginossar *et al*, 2007]. Available algorithms allow the prediction of miRNA targets: applied to the human cytomegalovirus miRNAs (hcmv-miRUL112), the top candidate target identified is the major histocompatibility complex class I–related chain B (MICB). The MICB is a stress-induced ligand of the natural killer (NK) cell activating receptor NKG2D and is critical for the NK cells killing virus-infected cells and tumor cells. The hcmv-miR-UL112 specifically down-regulates MICB expression during viral infection, leading to a decreased binding of NKG2D and reduced killing by NK cells.

This example reveals a miRNA-based immunoevasion mechanism that appears to be exploited by the human cytomegalovirus. This important result confirms the main impact the

regulatory networks exert on the expression of genes (like MICB, which is regulated by NU.1 on Figure 3bis top, violet arrows) involved in immune response regulation.

Considering the part of the network structure made of uptrees (coming for example from the gene Delta in the Figure 3, bottom), circuits, and downtrees (leading for example toward the different immuno-competent cell lineages, like NK lineage), the miRNAs may act as uptree sources. The frequency of the predicted miRNA binding sites in the immune genes 3′ UTRs indicated preferential targeting of these genes compared to the whole genome [Anderson *et al*, 2002; Zamoyska *et al*, 2003; Stern-Ginossar *et al*, 2007; Asirvathama *et al*, 2008]. Major targets include transcription factors, cofactors and chromatin modifiers whereas upstream factors, such as ligands and receptors (cytokines, chemokines and TLRs), are generally not targeted. About 10% of the immune genes were ‘hubs’ with eight or more different miRNAs predicted to target their 3′UTRs, which represents an important inhibitory "noise", leaving expressed only the positively regulated genes (by autocatalyses or by positive circuits as shown in Figure 3), even through an even number of inhibitions (like RAG in Figure 6, from GATA3, itself positively controlled by the MAP kinase on Figure 3 bis, inhibited by many human MiRNAs like 140-171, 173, 174, 176, 188, 306, 307, 313 and 321, from www.mirbase.org) like the acid leaves expressed the parts of an "aquaforte" non engraved in copper depth.

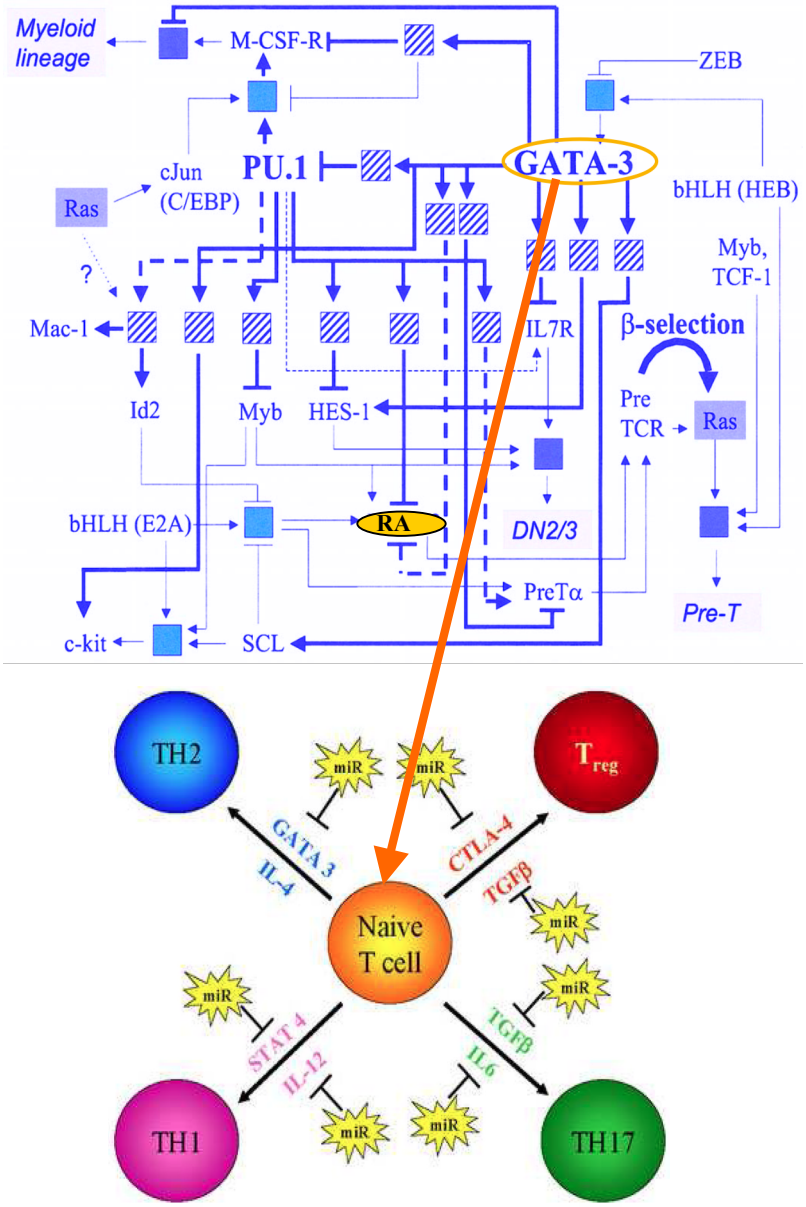


Figure 6. Regulation concerning the Naive T cell fate that involve microRNAs. After [Anderson *et al*, 2002].

In [Asirvathama *et al*, 2008], authors focused on certain key immune genes, such as BCLs (cf. Figure 3) and others presenting binding sites for miRNAs. The ubiquitory proteins NF-κB (a protein complex that controls the DNA transcrition and plays a key role in regulating the immune response) and p53 do not present such binding sites, but their pathways are targeted by miRNAs at downstream sites. MHC class II genes lacked miRNA targets but binding sites were identified in the CIITA gene and shown experimentally to repress IFN-γ induced MHC class II activation. Moreover, multiple components involved in the generation and effector

functions of miRNAs (Dicer and Argonautes) were themselves miRNA targets, suggesting that a subset of miRNAs may indirectly control their own production as well as other miRNAs, thus increasing the ubiquitory inhibitory noise caused by the small RNAs (si- and mi-RNAs). The Figure 7 illustrates the role of such small RNAs (like the CD4 silencer) in the regulation of T cell differentiation.

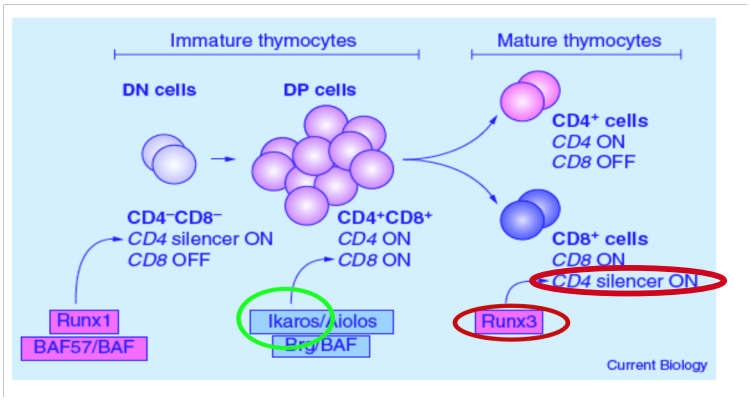


Figure 7. Regulated expression of CD4 and CD8 durin T cell differentiation, with identification of siRNA (like the CD4 silencer, in red). After [Zamoyska *et al*, 2003].

Chromatine dynamics

The ultimate regulation target of the immunetwork controlling antigenic loci rearrangements is the chromatin. As fully described in Chapter 1, in eukaryotes, linear DNA is associated with a protein complex of histones to constitute a compacted nucleoprotein complex, the nucleosome. The core of a nucleosome consists in 146 DNA base pairs wrapped around a histone octomer (each histone octomer containing two copies of four histones, H2A, H2B, H3, and H4). A linker DNA of roughly 50 bp long separates the nucleosome core particles. A supplementary H1 histone associates with the DNA of the nucleosomes and stabilizes the fibre. This structure corresponds to the 10-nm fibre, which further coils to form the 30-nm fibre.

Eukaryotic interphase chromatin includes heterochromatin domains highly condensed and mostly transcriptionally inactive, as well as less compacted euchromatin domains containing actively expressed genes. The denoted facultative heterochromatin corresponds to certain euchromatic areas and might be transcriptionally inactive or active depending on the cell lineage or developmental stage.

Chromatin remodeling and subnuclear relocalisations of the loci coding for the TR and Ig chains constitute prerequisites for the V(D)J rearrangements. The chromatin can be considered as

a mediated epigenetic system [Fossey *et al*, 2009], and its structural modifications would require distinctive time delays than gene-to-gene interaction ones within the immunetwork. Eventually, new updating rules should be developed, consisting in distinguishing the state of all genes involved in the chromatin dynamics ("chrodyn" genes). For example, if these particular genes are in the 0 state (no expression), then all the genes in the updating blocks depending on the state of these chrodyn genes, reached by the update dynamics after the shift of chrodyn genes to the state 0 (consecutive for example to a miRNA inhibition) would be put in the 0 state and no more updated until the chrodyn genes appear newly in state 1.This chromatin dynamics dependency would be taken into account by fixing large negative weights on the interaction arrows between the chrodyn genes and the genes depending on the chromatin activity to be updated. This may constitute open perspectives for further works in the continuity of the present thesis.

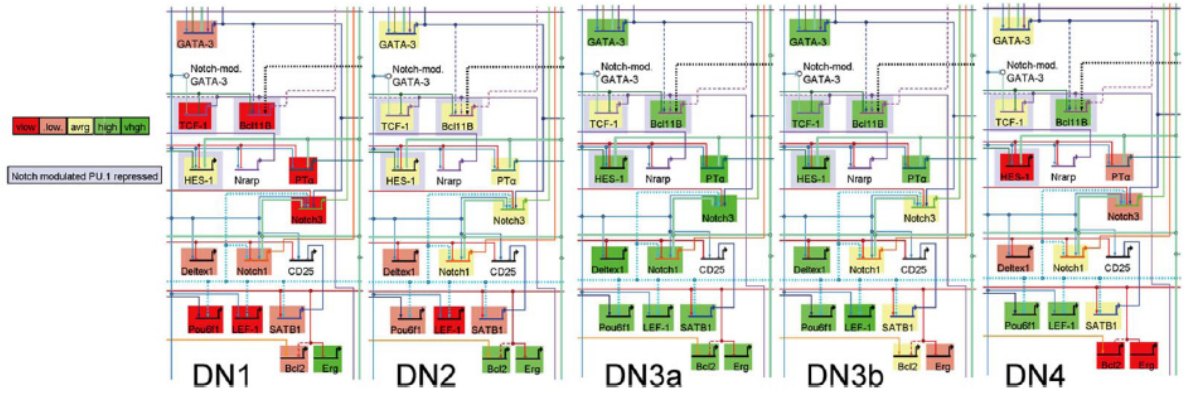
Mathematical inverse methods for immunetworks

The identification of the expression states of the genes involved in immunetworks through bio-arrays data or from bio-functional considerations allows, with the use of mathematical inverse methods, the clear deduction of the logic regulatory network, which generated the sequence of states. This extraction of a plausible and logic network architecture requires a precise knowledge about the end of the transient states of the data trajectories, namely the entrance time in the asymptotic regime of the stationary or cyclic attractors of the studied immunetwork.

For example, from the data presented by [Brink *et al*, 2009] and [Georgescu *et al*, 2008], reproduced on Figures 8 and 9, it is possible to deduce the expression state of important immune-related gens as PU.1, Notch1, Notch3, GATA3, Ikaros, Runx 1, Runx3, RAG 1, Bcl1, Lck, ZAP70, or LAT. These genes appear in the network presented on Figure 4, but some interactions are not strongly confirmed. The identified gene states, considered as asymptotic, permit to complete the immunetwork through inverse methods and afterwards to check if the hypothetic interactions were plausible or not [Aracena *et al*, 2004].

Name	Accession	Description	Ratio		
			Brain	Heart	Kidney
Igk-C	XM_132633	Immunoglobulin kappa chain, constant region	2.66	1.31	3.58
Slp	NM_011413	Sex-limited protein	2.12	2.48	1.89
Slp	NM_011413	Sex-limited protein	2.05	1.96	2.03
C4	NM_009780	Complement component 4 (within H-2S)	1.77	1.98	1.95
Cd52	NM_013706	CD52 antigen	1.72	1.28	1.44
Fcrl3	NM_144559	Fc receptor-like 3	1.68	1.75	1.68
C3	NM_009778	Complement component 3	1.63	1.42	1.39
Lyzs	NM_017372	Lysozyme	1.45	0.96	2.28
Bcl2a1a	NM_009742	B-cell leukemia/lymphoma 2 related protein A1a	1.39	1.70	1.94
Irak3	NM_028679	Interleukin-1 receptor-associated kinase 3	1.23	0.78	0.87
Igh-1a	XM_354704	Immunoglobulin heavy chain 1a (serum IgG2a)	1.13	3.89	7.33
Psmb8	NM_010724	Proteosome (prosome, macropain) subunit, beta type 8	1.04	0.86	0.81
Ii	NM_010545	Ia-associated invariant chain	0.99	1.16	0.95
Bcl2a1b	NM_007534	B-cell leukemia/lymphoma 2 related protein A1b	0.98	1.01	1.84
Rmcs1	NM_207105	Response to metastatic cancers 1	0.97	1.30	0.94
Serpina3n	NM_009252	Serine (or cysteine) proteinase inhibitor, clade A, member 3N	0.92	2.80	1.37
Ctss	NM_021281	Cathepsin S	0.86	0.79	1.88
Ms4a6d	NM_026835	Membrane-spanning 4-domains, subfamily A, member 11	0.86	0.84	3.15
Casp1	NM_009807	Caspase 1	0.81	0.78	1.16
Temt	NM_009349	Thioether *S*-methyltransferase	0.74	2.13	−2.22
Cd68	NM_009853	CD68 antigen	0.72	0.67	1.50
Ifi205	NM_172648	Interferon activated gene 205	0.70	0.96	0.64
Icam1	NM_010493	Intercellular adhesion molecule	0.68	0.98	0.61
Fcgr3	NM_010188	Fc receptor, IgG, low affinity III	0.65	0.61	1.30
C1qa	NM_007572	Complement component 1, q subcomponent, alpha polypeptide	0.64	0.73	1.18
C1qg	NM_007574	Complement component 1, q subcomponent, gamma polypeptide	0.64	0.62	1.16
Tcrb-V8.2	NC_000072	T-cell receptor beta, variable 8.2	−0.66	0.99	1.41
Itm2a	NM_008409	Integral membrane protein 2A	−0.66	−0.79	−0.73

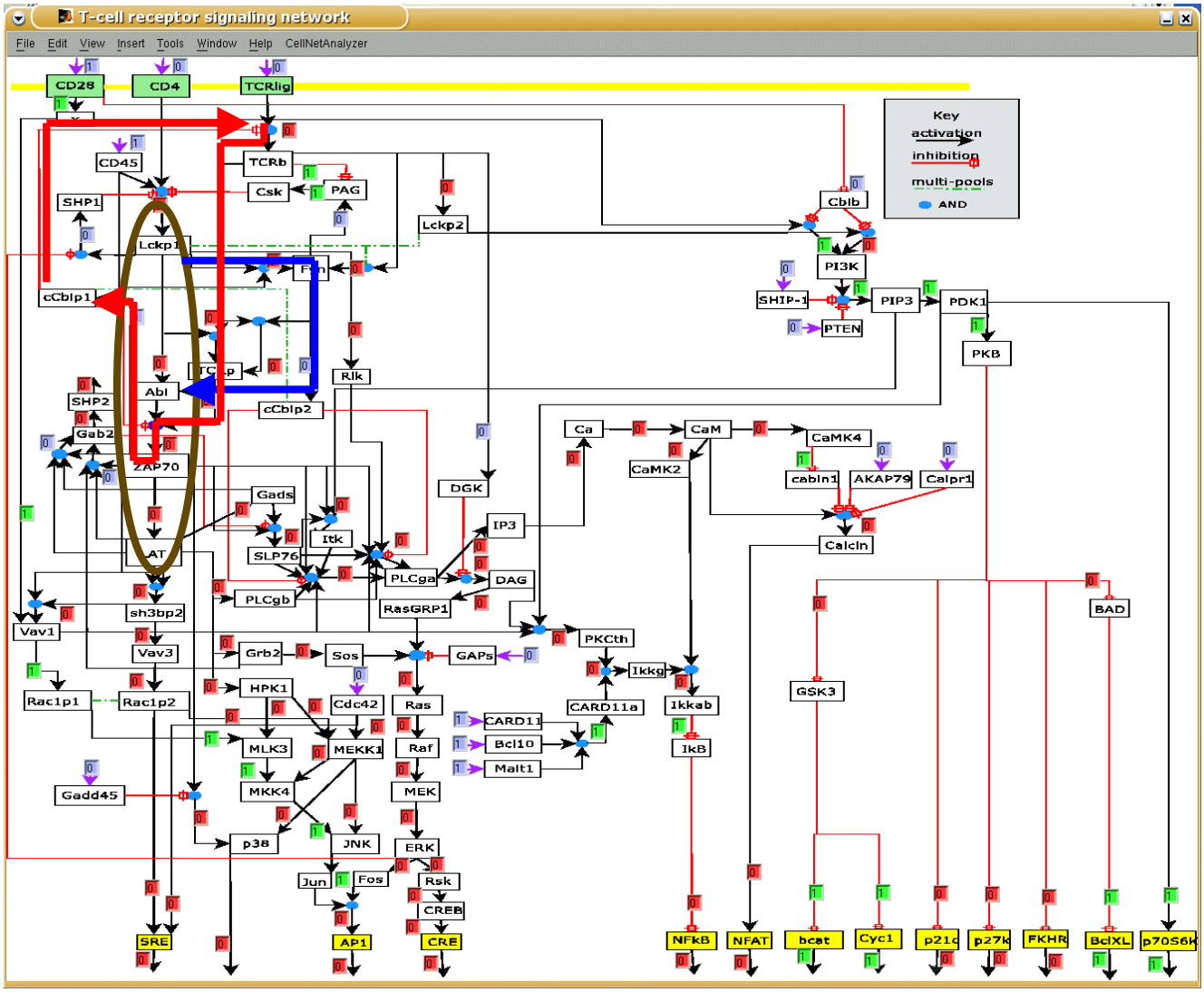
Figure 8. Overexpression of immune-related genes in ageing. After [Brink *et al*, 2009].



Gene category	Initiation, stem to DN1 increase	DN2 entry, DN1 to DN2 increase	Specification, DN1–DN3 continuous		Commitment, DN3 stage		β-Selection	
			Increase	**Decrease**	Increase	**Decrease**	Increase	**Decrease**
Transcription factors	GATA-3 TCF-1 HES1	Bcl11b HEB-alt	GATA-3 TCF-1 (Tcf7) HEB (Tcf12) Ikaros Gfi1 Runx1 Ets1	**PU.1** **SCL** **GATA-2** **Id2** **Runx2** **Runx3** **Gfi1B** **C/EBPα** **(& others)**	Ets2 Pou6f1 SpiB HES1 LEF-1 Id3	**PU.1** **SCL**	Aiolos SATB1 Tbet	**SpiB** **HES1** **Myb** **Runx1** **HEB-alt** **Erg** **Id2** **(and others)**
Targets/other	Deltex1	CD3γ CD3ε	IL-7Rα CD25 Rag-1 CD3ε Lck LAT ZAP70	**Mac-1**	Deltex1 pTα (Ptcra) Rag-1 Notch3 Lck LAT ZAP70	**c-Kit**	TCR-Cα ZAP70	**Deltex1** **IL-7Rα** **pTα (Ptcra)** **Rag-1** **CD25** **Bcl2** **Notch1**

Figure 9. Close-ups of one region of the network with background highlighting indicating differential gene expression levels at five different developmental states (top). Gene expression changes marking transitions in T-cell development (bottom). After [Georgescu *et al*, 2008].

To illustrate this method, the network presented on Figure 10 from [Saez-Rodriguez *et al*, 2007] brings the possibility to check the plausible and logic character of the network which was located downstream the genes ZAP 70 – LAT of the Figures 3 bis and 11 (marron circle). An approach based on constraint programming can be useful for the explicitation of all networks compatible with these states considered as attractors [Ben Amor *et al*, submitted]. A way to help this identification is to calculate (from the table of Figure 10) the number of the attractors corresponding to the two intersecting circuits, negative of size 5 (red) and positive of size 4 (blue) giving 3 attractors, the half of the number observed for an isolated positive circuit of lengh 4 (violet on Figure 11). In case of intersecting circuits of Figure 3, the gain is from 5 to 20 (blue on Figure 11).



Input/ Output		WT	WT	WT	PI3K	PI3K	PI3K	SLP76	Fyn	Fyn	Fyn	Rlk and Itk	Lck and Fyn	Lck and Fyn	Lck and Fyn
Input	TCR	1	0	1	1	0	1	1	1	1	1	1	1	0	1
	CD4	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	CD28	0	1	1	0	1	1	0	0	0	1	0	0	1	1
Output	ZAP	1	0	1	1	0	1	1	0	1	0	1	0	0	0
	LAT	1	0	1	1	0	1	1	0	1	0	1	0	0	0
	PLCga	1	0	1	0	0	0	0	0	1	0	0	0	0	0
	ERK	1	0	1	0	0	0	0	0	1	0	0	0	0	0
	JNK	1	1	1	1	1	1	1	0	1	1	1	0	1	1
	PKB	1	1	1	0	0	0	1	1	1	1	1	0	1	1
	AP1	1	0	1	0	0	0	0	0	1	0	0	0	0	0
	NFKB	1	0	1	0	0	0	0	0	1	0	0	0	0	0
	NFAT	1	0	1	0	0	0	0	0	1	0	0	0	0	0
	Reference	Figure 3A, 3C	Figure 3A, 3D	Figures 3A, 4	Figure 3C	Figure 3D	Figure 4	[49]	Figure 3B, [50]	Figure 3B, [50]	Figure 3B, [50]	[51]	Figure 4	Figure 4	Figure 4

Figure 10. Immunetwork downstream the genes ZAP 70 – LAT. After [Saez-Rodriguez *et al*, 2007].

The last example concerns the regulation of the T helper cells maturation. The immunetwork on Figure 11 exhibits one negative circuit of length 8 (blue and red) intersecting a positive circuit of length 5 (red), and also one negative circuit of length 3 (encircled in violet) intersecting the positive circuit of length 5. From the Figure 12 (violet circle on the top table), we deduce that the immunetwork displays 3 attractors from these last intersecting circuits, which is less than the 8 attractors of an isolated positive circuit of length 5 (green circle on the bottom table). This last example is another proof that the reduction of the attractors number, i.e., the evolution toward "pauci-functional" networks (controlling only few dedicated functions) could be one of the characteristics of the immune system.

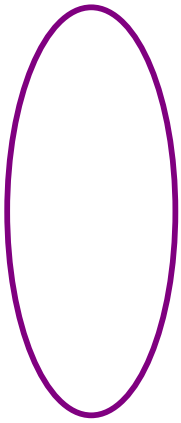
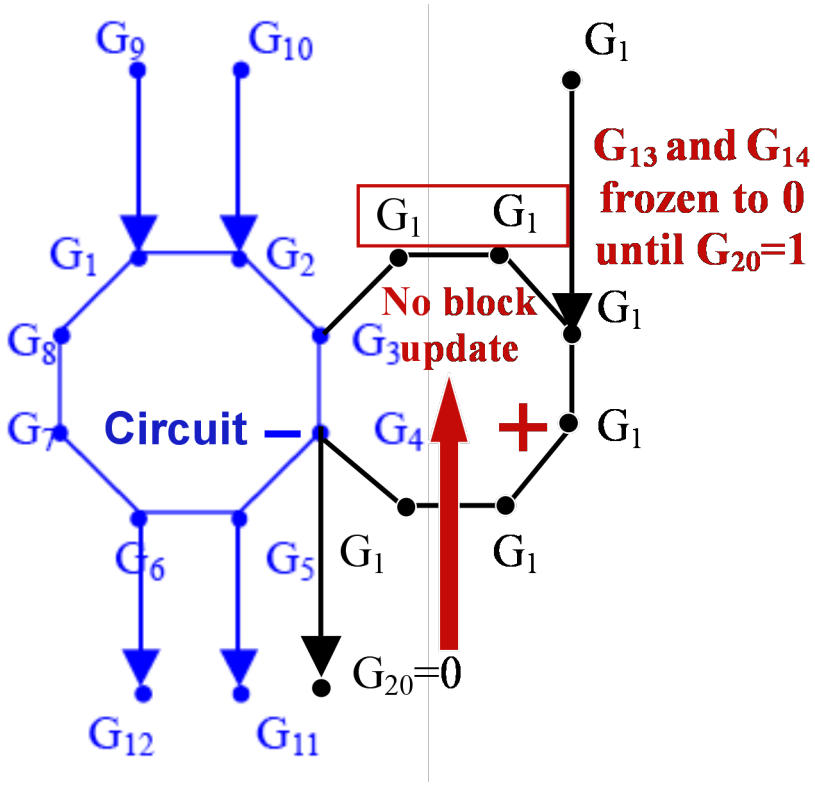


Figure 11. Immunetwork upstream the gene GATA3. After [Mendoza, 2006].

Further studies about immunetworks will be performed in a next future in order to elucidate the exact role of the intersecting circuits. In fact, the immunetworks would constitute a dedicated application example for future theoretical discoveries [Demongeot, Elena *et al*, submitted].

A



B

ℓ \ r	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	2	2	3	3	5	5	8	10	15	19	31	41	64
2	1	1	2	3	3	4	5	8	10	14	19	31	41	63
3	1	2	1	3	3	6	5	8	8	15	19	33	41	64
4	1	1	2	1	3	4	5	11	10	14	19	24	41	63
5	1	2	2	3	1	5	5	8	10	26	19	31	41	64
6	1	1	1	3	3	1	5	8	8	14	19	63	41	63
7	1	2	2	3	3	5	1	8	10	15	19	31	41	158
8	1	1	2	1	3	4	5	1	10	14	19	24	41	63
9	1	2	1	3	3	6	5	8	1	15	19	33	41	64
10	1	1	2	3	1	4	5	8	10	1	19	31	41	63
11	1	2	2	3	3	5	5	8	10	15	1	31	41	64
12	1	1	1	1	3	1	5	11	8	14	19	1	41	63
13	1	2	2	3	3	5	5	8	10	15	19	31	1	64
14	1	1	2	3	3	4	1	8	10	14	19	31	41	1

C

p \ n	1	2	3	4	5	6	7	8	9	10	11	12	21	22
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	–	1	–	1	–	1	–	1	–	1	–	1	–	1
3	–	–	2	–	–	2	–	–	2	–	–	2	2	–
4	–	–	–	3	–	–	–	3	–	–	–	3	–	–
5	–	–	–	–	6	–	–	–	–	6	–	–	–	–
6	–	–	–	–	–	9	–	–	–	–	–	9	–	–
7	–	–	–	–	–	–	18	–	–	–	–	–	18	–
8	–	–	–	–	–	–	–	30	–	–	–	–	–	–
9	–	–	–	–	–	–	–	–	56	–	–	–	–	–
10	–	–	–	–	–	–	–	–	–	99	–	–	–	–
11	–	–	–	–	–	–	–	–	–	–	186	–	–	186
12	–	–	–	–	–	–	–	–	–	–	–	335	–	–
21	–	–	–	–	–	–	–	–	–	–	–	–	99858	–
22	–	–	–	–	–	–	–	–	–	–	–	–	–	190557
T_n^+	2	3	4	6	8	14	20	36	60	108	188	352	99880	190746

Figure 12. A. Generic circuitry showing the presence intersecting circuits and calculation of the attractor number due to these intersecting circuits in case of (**B.**) negative/positive circuits intersection and of (**C**.) a unique positive circuit. After [Demongeot, Noual *et al*, submitted].

p \ n	1	2	3	4	5	6	7	8	15	16	17	18	21	22
2	1	–	1	–	1	–	1	–	1	–	1	–	1	–
4	–	1	–	–	–	1	–	–	–	–	–	1	–	1
6	–	–	1	–	–	–	–	–	1	–	–	–	1	–
8	–	–	–	2	–	–	–	–	–	–	–	–	–	–
10	–	–	–	–	3	–	–	–	3	–	–	–	–	–
12	–	–	–	–	–	5	–	–	–	–	–	5	–	–
14	–	–	–	–	–	–	9	–	–	–	–	–	9	–
16	–	–	–	–	–	–	–	16	–	–	–	–	–	–
30	–	–	–	–	–	–	–	–	1091	–	–	–	–	–
32	–	–	–	–	–	–	–	–	–	2048	–	–	–	–
34	–	–	–	–	–	–	–	–	–	–	3855	–	–	–
36	–	–	–	–	–	–	–	–	–	–	–	7280	–	–
42	–	–	–	–	–	–	–	–	–	–	–	–	49929	–
44	–	–	–	–	–	–	–	–	–	–	–	–	–	95325
T_n^-	1	1	2	2	4	6	10	16	1096	2048	3856	7286	49940	95326

Figure 13. Calculation of the attractor number due to a unique negative circuit. After [Demongeot, Noual *et al*, submitted].

Immunetworks and ageing

Generally, the ageing process can suppress the functional "canalization", leading to the presence of an unique attractor for the immunetworks, e.g., by favoring the domination of the negative circuit, through, for example, the over-expresson of the gene Irak3 (Figure 9, orange circle). Another effect of ageing on the immune system is given in [Braeckman *et al*, 2006], where the assumption is made that the anti-aging effect of the dietary restriction (DR) implicates that signalling pathways should exist that link nutrient sensing with an appropriate effector mechanism to enhance somatic maintenance. One candidate pathway is the Insulin/IGF-1 signalling pathway. A good deal of the mutations that extend life span in *C. elegans* defines genes that are homologues of mammalian Ins/IGF-1 signaling pathway components. Although, at first glance, it might seem plausible that the Ins/IGF-1 pathway mediates the anti-aging effect of DR, there are several strong arguments against this assumption. Most of these are based on the additive effect of DR and reduced Ins/IGF-1 signalling. Inflammation is produced by eicosanoids and cytokines, which are released by injured or infected cells. Common cytokines include interleukins that are responsible for communication between white blood cells, chemokines that promote chemotaxis, and interferons that have anti-viral effects [Brink *et al*, 2009]. In the list of overlapping age-related genes, at least five genes are present that are involved in the inflammatory response - Casp1, Irak3 (involved in the regulation of the T helper cells maturation, cf. Figure 12), Cd48, Dock2, and Icam1. Casp1 was identified by its ability to proteolytically cleave and activate the inactive precursor of interleukin-1. The expression of its human homologue and the subsequent release of IL-1beta and IL-18 (also involved in the regulation of the T helper cells maturation, cf. Figure 12) significantly contribute for instance to intestinal inflammation. Moreover, it was recently concluded that IL-1beta and IL-18 participate in fundamental inflammatory processes that increases during aging [Dinarello *et al*, 2006]. Also directly related to IL-1 are the interleukin-1 receptor associate kinases, a member, Irak3, regulates innate immunity through unknown mechanism [Su *et al*, 2007]. Cd48 is an interleukin (IL)-3-induced activating receptor on eosinophils, which may be involved in promoting allergic inflammation [Munitz *et al*, 2006]. Dock2 is a member of chemokines that promote chemotaxis, which has been shown to be of key importance for lymphocyte chemotaxis [Fukui *et al*, 2001]. Most interestingly, Icam1 (cf. Figure 9, red cycle) is one of the proteins involved in inflammatory responses, and it is overexpressed in senescent cells and aged tissues. Additionally, the NF-kB signaling cascade is crucial for the activation of human Icam1 expression in response

to inflammation [Kletsas *et al*, 2004].

More generally, it would be of great interest to elucidate the role of the immune system on ageing process acceleration either through the general inflammatory response or through more specific auto-immune pathologies, with down- or over-expressed genes involved in the immune system regulation. This would constitute a future perspective of our present work.

References of the Chapter IV

Anderson MK, Hernandez-Hoyos G, Dionne CJ, Arias AM, Chen D, Rothenberg EV. Definition of Regulatory Network Elements for T Cell Development by Perturbation Analysis with PU.1 and GATA-3. Developmental Biology. 2002. 246:103-121.

Aracena J, Demongeot J. Mathematical Methods for Inferring Regulatory Networks Interactions: Application to Genetic Regulation. Acta Biotheoretica. 2004. 52:391-400.

Asirvathama AJ, Gregoriea CJ, Hub Z, Magnera WJ, Tomasia TB. MicroRNA Targets in Immune Genes and the Dicer/Argonaute and ARE Machinery Components, Mol. Immunol. 2008. 45:1995-2006.

Ben Amor H, Corblin F, Fanchon E, Elena A, Trilling L, Demongeot J, Glade N. Formal Methods for Hopfield-like networks. PNAS. submitted.

Berge C. Graphes et Hypergraphes, Paris : Dunod. 1974.

Bollobas B. Random Graphs, London : Academic Press. 1985.

Braeckman BP, Demetrius L, Vanfleteren JR. The dietary restriction effect in *C. elegans* and humans: is the worm a one-millimeter human? Biogerontology. 2006. 7:127-133.

Brink TC, Regenbrecht C, Demetrius L, Lehrach H, Adjaye J. Activation of the immune response is a key feature of aging in mice. Biogerontology. 2009. 10:721-734.

Cinquin O, Demongeot J. Positive and negative feedback : mending the ways of sloppy systems. C.R. Acad. Sci. Biologies. 2002. 325:1085-1095.

Cinquin O, Demongeot J. Positive and negative feedback : striking a balance between necessary antagonists. J. Theoret. Biol. 2002. 216:239-246.

Delbrück M. Discussion, in: Unités biologiques douées de continuité génétique, Colloques Internationaux CNRS. 1949. 833-35.

Demongeot J, Noual M, Sené S. Combinatorics of Boolean automata circuits dynamics. Discrete Applied Mathematics. submitted.

Demongeot J, Elena A, Noual M, Sené S, Taramasco C, Thuderoz F. Attractors of intersecting circuits and applications to "immunetworks". J. Theor. Biol. submitted.

Demongeot J, Ben Amor H, Gillois P, Noual M, Sené S. Robustness of regulatory networks. A Generic Approach with Applications at Different Levels: Physiologic, Metabolic and Genetic. Int. J. Molecular Sciences. 2009. 10:4437-4473.

Demongeot J, Thuderoz F, Baum TP, Berger F, Cohen O. Bio-array images processing and genetic networks modeling. C. R. Acad. Sci. Biologies. 2003. 326:487-500.

Demongeot J, Aracena J, Thuderoz F, Baum TP, Cohen O. Genetic regulation networks: circuits, regulons and attractors. C. R. Acad. Sci. Biologies. 2003. 326:171-188.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modeling. ISBI. 2002. IEEE EMB, M. Unser & Z.P. Liang, Eds., Piscataway : IEEE Press, 50-54 2002.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modeling. in : Modelling & simulation of biological processes in the context of genomics, P. Amar et al., Eds., Evry : Genopole. 2002. 87-94.

Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. Bio-array images processing and genetic networks modeling. 5th IEEE EMBS International Summer School on Medical Imaging, J.L. Coatrieux et al., Eds., Piscataway : IEEE Press, 15-23, 2002.

Demongeot J. Multi-stationarity and cell differentiation. J. Biol. Syst. 1988. 6:1-2.

Dinarello CA. Interleukin 1 and interleukin 18 as mediators of inflammation and the aging process. Am. J. Clin. Nutr. 2006. 83:447–455.

Fossey A. Epigenetics: Beyond genes, Southern Forests. Journal of Forest Science. 2009. 71: 121-124.

Georgescu C, Longabaugh WJR, Scripture-Adams DD, David-Fung ES, Yui MA, Zarnegar MA, Bolouri H, Rothenberg EV. A gene regulatory network armature for T lymphocyte specification. PNAS. 2008. 105:20100–20105.

Fukui Y, Hashimoto O, Sanui T, Oono T, Koga H, Abe M, Inayoshi A, Noda M, Oike M, Shirai T, Sasazuki T. Haematopoietic cell-specific CDM family protein DOCK2 is essential for lymphocyte migration. Nature. 2001. 412:826–831.

Hennecke J, Wiley DC. T cell receptor-MHC interactions up close. Cell. 2001. 104 :1-4.

Kauffman S. The Origins of Order. Oxford (UK) : Oxford University Press. 1993.

Kletsas D, Pratsinis H, Mariatos G, Zacharatos P, Gorgoulis VG. The proinflammatory phenotype of senescent cells: the p53-mediated ICAM-1 expression. Ann. N. Y. Acad. Sci. 2004.1019:330–332.

Mendoza L. A network model for the control of the differentiation process in Th cells. BioSystems. 2006. 84:101-114.

Mendoza L, Alvarez-Buylla ER. Dynamics of the genetic regulatory network for *Arabidopsis* thaliana flower morphogenesis. J. Theoret. Biol. 1998. 193:307-319.

Munitz A, Bachelet I, Eliashar R, Khodoun M, Finkelman FD, Rothenberg ME, Levi-Schaffer F. CD48 is an allergen and IL-3-induced activation molecule on eosinophils. J. Immunol. 2006. 177:77–83.

Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, Haus UU, Weismantel R, Gilles ED, Klamt S, Schraven B. A Logical Model Provides Insights into T Cell Receptor Signaling. PLoS Comput. Biol. 2007. 3-e163.

Singh H, Grossched R. Molecular analysis of B lymphocyte development and activation, Heidelberg : Springer Verlag, 2005.

Snoussi EH. Necessary condition for multi-stationarity and stable periodicity. J. Biol. Syst. 1998. 6:3-10.

Stern-Ginossar N, Elefant N, Zimmermann A, Wolf DG, Saleh N, Biton M, Horwitz E, Prokocimer Z, Prichard M, Hahn G, Goldman-Wohl D, Greenfield C, Yagel S, Hengel H, Altuvia Y, Margalit H, Mandelboim O. Host Immune System Gene Targeting by a Viral miRNA. Science. 2007. 317:376-380.

Su J, Xie Q, Wilson I, Li L. Differential regulation and role of interleukin-1 receptor associated kinase-M in innate immunity signaling. Cell Signal. 2007. 19:1596–1601.

Thomas R. On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. Springer Series in Synergetics. 1980. 9:1-23.

Zamoyska R. T-cell differentiation: chromatin remodelling in CD4/CD8 regulation. Curr. Biol. 2003. 13:R189-191.

Annex 1

Bio-array images processing and genetic networks modeling. Demongeot J, Berger F, Baum TP, Thuderoz F, Cohen O. ISBI. 2002. IEEE EMB, M. Unser & Z.P. Liang, Eds., Piscataway : IEEE Press, 50-53.

This paper appears in: Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on Biomedical Imaging
Issue Date: 2002 **On page(s):** 50 - 53
Print ISBN: 0-7803-7584-X **INSPEC Accession Number:** 7516427
Digital Object Identifier: 10.1109/ISBI.2002.1029190 **Date of Current Version:** 07 novembre 2002

BIO-ARRAY IMAGES PROCESSING AND GENETIC NETWORKS MODELLING

J. Demongeot, F. Berger, T.P. Baum, F. Thuderoz & O. Cohen*

TIMC-IMAG CNRS 5525 Faculty of Medicine 38 700 La Tronche France
*INSERM U 318 University Hospital of Grenoble 38 700 La Tronche France

ABSTRACT

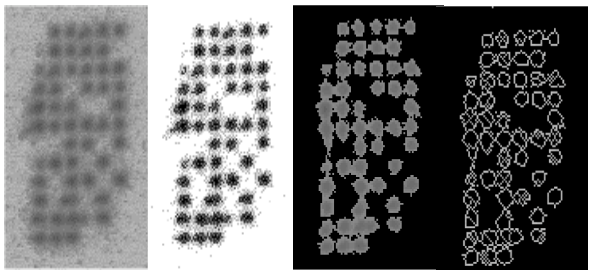
The new tools available for gene expression studies are essentially the bio-array methods using a large variety of physical detectors (isotopes, fluorescent markers, ultrasounds,…). Here we present an image processing method independent of the detector type, dealing with the noise and with the peaks overlapping, the peaks revealing the detector activity (isotopic in the presented example), correlated with the gene expression. After this first step of image processing, we can extract information about causal influence (activation or inhibition) a gene can exert on other genes, leading to clusters of co-expression in which we extract an interaction matrix explaining the dynamics of co-expression correlated to the studied tissue function.

1. INTRODUCTION

The total mRNA's of genes to test are extracted from the studied tissue (in the present case a glioma tissue). DNA's are synthesized by reverse transcription from these mRNA's including bases labeled with the isotope P^{33}. Resulting DNA's are then tested against identified complementary DNA's (cDNA targets), previously amplified by PCR and fixed on a nylon gel. The hybridization results are revealed in a phospho-imager and yield a digital image coming from the radioactive hybridization plate, called the bio-array image or shortly the bio-image. cDNA hybridized with a P^{33}DNA means that the complementary sequence of the P^{33}DNA is present in the related mRNA proving that the corresponding gene is expressed in the studied tissue.

2. PEAK SEGMENTATION

The first encountered problem is the fact that the bio-images are extremely noisy and that we have to low-pass them in order to suppress the high-frequency noise (see Figure 1). The result of this pre-treatment is a better separation of the isotopic activity peaks, allowing a watershed separation and contouring [1]. Then we will apply a more accurate segmentation and contouring method called the potential-hamiltonian or "Gaussian stamping" method: let us remark that the peaks are about Gaussian, with a relatively weak kurtosis and skewness allowing in particular the respect of the conservation "law": 2/3 of the peak activity are concentrated into the set of points (x,y) where the Gaussian curvature H(x,y) vanishes, i.e. inside the maximum gradient line of the peak. By exploiting this property, it is possible to neglect the part of the peak outside the projection of this remarkable line, called in the following the characteristic line, its equation being:
$H(x,y)=\partial^2 g/\partial x^2 \partial^2 g/\partial y^2 - (\partial^2 g/\partial x \partial y)^2 = 0$, where g(x,y) is the gray function at the pixel of coordinates (x,y).



a) b) c) d)

Figure 1: a) raw data b) low pass filtering c) watershed segmenting and d) contouring

We are thus led to consider the new gray function H(x,y) instead of the function g(x,y) and its level line H(x,y)=0. We display after a plane differential system of which the characteristic line is a limit cycle. Let H'(x,y) be the function defined by: $H'(x,y)=|H(x,y)|$. Vanishing of H'(x,y) occurs on the characteristic line (see Figure 2 for the visualization of g and H') and if we consider the following crude system:
$dx/dt = -\alpha \partial H'/\partial x + \beta \partial H'/\partial y$, $dy/dt = -\alpha \partial H'/\partial y - \beta \partial H'/\partial x$,
where α and β are real parameters, then the first part of this differential system is of steepest descent potential nature and along this flow, the orbits converge to the set of zeros of H'(x,y), on which the second part of convective Hamiltonian type becomes preponderant [2]. Parameters α and β are used to tune the speed of convergence to the limit cycle. To cope with random noise and numeric instabilities, we modify slightly the system into:

$dx/dt=-\alpha\partial H'/\partial x[H(x,y)/G(x,y)]+\beta\partial H'/\partial y$,
$dy/dt= -\alpha\partial H'/\partial y[H(x,y)/G(x,y)]-\beta\partial H'/\partial x$,
where $G(x,y)=||\text{grad}(g)||^2$.

g Gaussian peak G

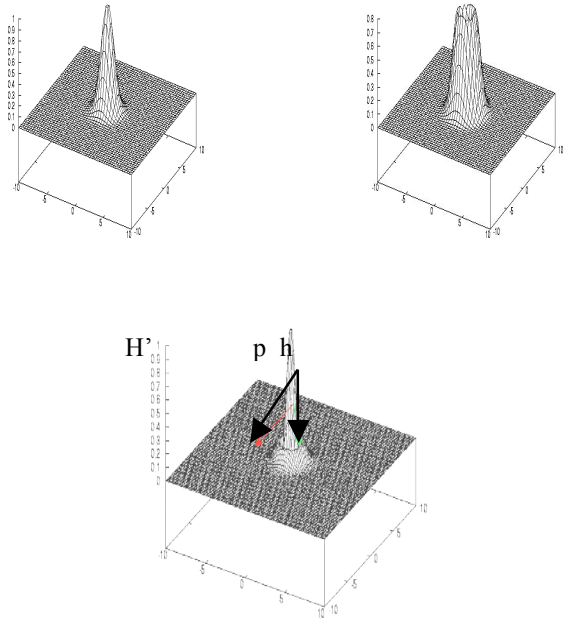


Figure 2: representation of g, G and H' (with indication of potential p and Hamiltonian h parts) for a Gaussian peak

The added term [H(x,y)/G(x,y)] speeds up the descent to the vanishing of H(x,y) and forces the stability. The usual discretization of Runge-Kutta yields ultimately the algorithm which is quite easy to implement. On each pixel (i,j) (boundary effects being neglected), the function H(i,j) reads:

$H(i,j)=[g(i+2,j)-2g(i+1,j)+g(i,j)][g(i,j+2)-2g(i,j+1)+g(i,j)]-[g(i+1,j+1)-g(i,j+1)-g(i+1,j)+g(i,j)]^2$.

We have seen that an important property of the characteristic line was that in the case of a Gaussian peak, it delimits a volume equal to 2/3 of the total volume of the peak. This property remains about exact in case of kurtosis and skewness of the peak. Hence by multiplying by 3/2 this volume, we get a good estimation of the gene activity and this value is better than those obtained by a watershed method due to over-segmentation (Figure 1). This approach is interesting because the lower part of the peak is often noisy. The method seems particularly efficient when the mesas are well separated. If they are close (Figure 3), then we need to tune the parameters α and β (Figure 4). In further developments of the method, we look for a dynamical calculation of these parameters from the data. Finally, we can standardize the estimated activity in terms of a bio-image with small squares symbolizing in gray levels the degree of hybridization of the cDNA's (Figure 5). From such bio-images acquired at different times of the cell cycle in cells from the same tissue, we can study the interactions between genes by estimating an interaction matrix.

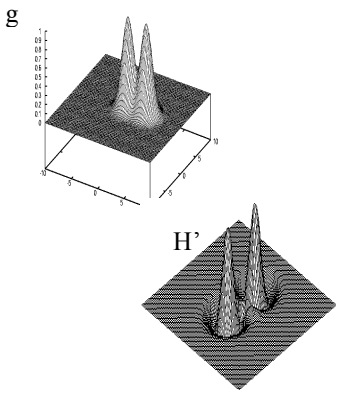


Figure 3: g (top) and H' (bottom) for close Gaussian peaks

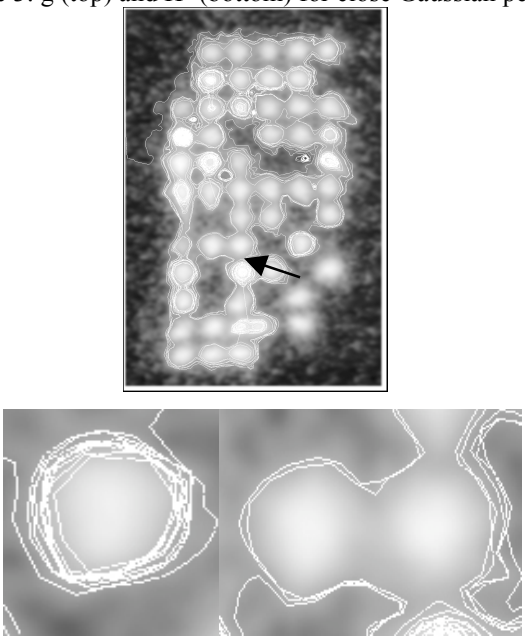


Figure 4: treated bio-image (top), succeeding limit cycle (left bottom) and false contour (↖ and right bottom)

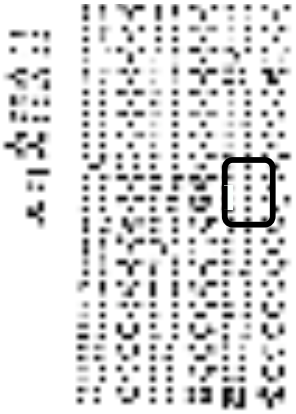


Figure 5: standardized bio-image

3. INTERACTION MATRIX

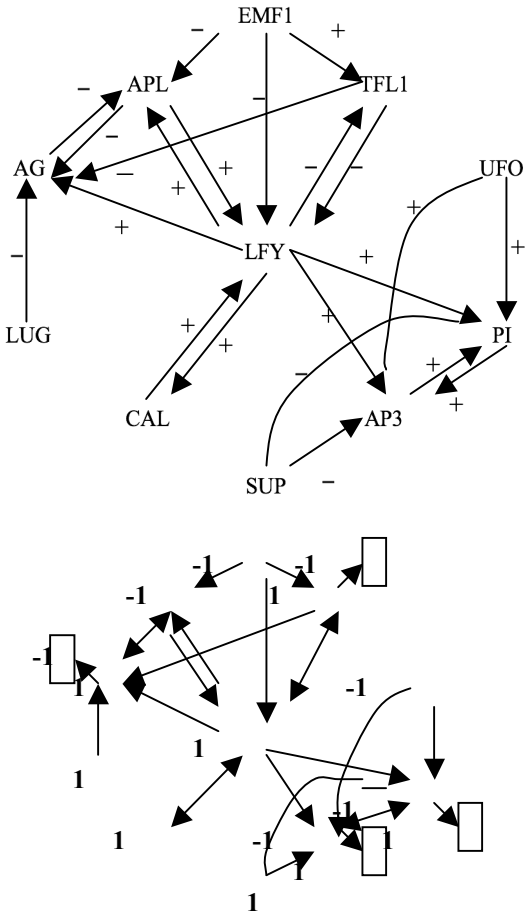


Figure 6: interaction graph of the flowering operon of *Arabidopsis* thaliana (top) and an attractor of its Boolean dynamics (bottom)

For each operon, we can define an interaction matrix M, which just expresses that if its coefficient m_{ij} is positive (resp. negative), the gene j is a promoter or activator (resp. repressor or inhibitor) of the gene i. If m_{ij} is null, then the gene j has no influence on the expression of the gene i. The interaction graph can be built from the interaction matrix M (Figure 6) by drawing an edge + (resp. -) between the vertices representing the genes j and i, iff $m_{ij}>0$ (resp. <0). In order to calculate the m_{ij}'s, we can either determine the s-directional correlation $\rho_{ij}(s)$ between the state vector $\{x_j(t-s)\}_{t\in C, t\geq s}$ of gene j at times t-s and the state vector $\{x_i(t)\}_{t\in C, t\geq s}$ of gene i at times t, t varying during the cell cycle C, or identify the system with a Boolean neural network. We define the connectivity K(M) of the interaction matrix M by the ratio between the numbers m of edges of the interaction graph and n of vertices: in general, for known operons (lactose operon, Cro operon of the phage λ, lysogenic/lytic operon of the phage μ infecting E. coli, gastrulation operon,…), K(M) is between 1.5 and 3. The observed induction proportion (number of positive edges divided by m) is between 1/3 and 1/2. If m_{ij}'s are unknown, we can take them randomly by respecting connectivity and induction proportion.

4. GENETIC NETWORK DYNAMICS

If we consider the interaction graph of the flowering operon of *Arabidopsis* thaliana (Figure 6 top) [3], then we can easily define from it a Boolean dynamics with threshold 0: the gene i has the state 1 if it is expressed and -1 if not. The change of state of gene i between the times t and t+1 obeys a majority rule, i.e. we calculate the numbers of its neighbors in state 1 with positive interaction and with negative interaction: if these 2 numbers are equal, then the new state of i is 1; if the activatory (resp. inhibitory) neighbors dominate, then the new state of i is 1 (resp. -1). When the time t is increasing, the configuration of genes states reaches a stable set of configuration (either a fixed configuration or a cycle of configurations) called an attractor of the genetic network dynamics. In Figure 6 (bottom), an example of such an attractor is given, with final states (in black boxes) different from the initial conditions.

We will present now first some qualitative results from the human genome observation, and after some theoretical corresponding statements recently proved :

- in 1948, M. Delbrück [4] conjectured that the presence of positive loops (i.e. paths from a gene i to itself having an even number of inhibitions [5]) in the interaction graph was a necessary condition for the cell differentiation; this conjecture has been more precisely written in a good mathematical context by R. Thomas in 1980 [6]
- in 1992, S. Kauffman [7] conjectured that the mean number of attractors for a Boolean genetic network with n genes and with connectivity 2, was equal to $\sqrt{n}$. This conjecture is supported by real observations: we have about 35 000 genes in the human genome and about 200 different tissues, which can be considered as different attractors of the same dynamics. For *Arabidopsis* thaliana, K(M)=22/11=2 and there is $4\approx\sqrt{11}$ different tissues (sepals, petals, stamens, carpels) [3] and for the Cro operon [8] of the phage λ, K(M)=14/5=2.8 and there is $2\approx\sqrt{5}$ observed (lytic and lysogenic) attractors.

Recently [9-16] these conjectures have been partially proved:

Proposition 1: if all loops of the interaction graph are positive, then there exists a state vector $x = (x_1 \ldots, x_n)$ in $\{-1,1\}^n$, such that x and $-x = (-x_1, \ldots, -x_n)$ are fixed configurations of the genetic network dynamics.

Proposition 2: if all loops of the interaction graph are negative, then there is no fixed configuration.

Proposition 3: let a genetic network having n genes and n interactions, then a necessary and sufficient condition of existence of a fixed configuration x is the existence of a positive loop and –x is also a fixed configuration.

Proposition 4: given a state vector x, the set of minimal matrices M having x as fixed configuration is given by the following conditions:

1) $m_{ij} = a_{ij} x_i x_j$, where $a_{ij} \geq 0$ and, for all i, there exists j(i) such that $a_{ij(i)} > 0$
2) $-|a_{ij(i)}| < b_i \leq |a_{ij(i)}|$, where a_{ij}'s & bi's are weights and thresholds of the corresponding genetic network [14].

Proposition 5: if m is the total number of positive loops C, then the number of fixed configurations is less or equal than 2^m, and this upper bound is reached, if and only if for any positive loop C, there is no edge $(x,y)/x \notin C, y \in C$.

Proposition 6: if the genetic network has n genes and 2n interactions, then the expectation of the number of its fixed configurations is $\sqrt{n}$, if n is sufficiently large [15].

5. OPEN PROBLEMS

An important open problem concerns the relationship between the number F of fixed configurations and the number S of interaction loops of the interaction matrix M: the problem is in fact to find the best upper bound for F for a given interaction matrix M. This question is the discrete translation of the famous XVIth Hilbert's problem of determining an efficient upper bound for the number of limit cycles of a polynomial differential system. Let us summarize the role of the architecture of positive and negative (with an odd number of inhibitions) loops of M on the occurrence of multiple stationary behaviors as obtained above: if the number of genes and the number of interactions are the same, there is only one isolated loop in M and either this loop is negative and the lowest bound (0) for F is reached, or this loop is positive and the upper bound (2^1) for F is reached. If the numbers of genes and interactions are respectively n and n+1, there is 2 interaction loops with the following structure: if both loops are negative, F = 0; if there is a positive loop and a negative loop disjoint, F = 0; if there is a positive loop intersecting a negative loop, F = 1; if there is a positive loop intersecting a positive loop, F = 1; if there is two disjoint positive loops, F = 2^2. If more generally the number S of loops is m, then: if all loops are negative, F = 0; if all loops are positive, then: $2 \leq F \leq 2^m$, and if and only if all loops are positive and disjoint, F = 2^m. An interesting open problem is now to make exhaustive the determination of F and S and in particular to find the circumstances (related to the loops structure) in which we can relate the number of intersecting and isolated loops to F. The approach for solving this open problem could consist first in finding coherent relationships between analogous properties discovered for continuous versions of the regulatory networks and for general Boolean networks.

6. CONCLUSION

A geneticist could exploit the results above in the following sense: we have shown in the paper that it would be possible to characterize the minimal interaction matrices having certain state vectors as fixed configurations. The determination of these matrices is not unique, but permits to focus on certain important equivalence classes in which the expected matrix has to belong. This considerably restricts the choice of the possible interaction matrices compatible with observed fixed configurations, when it is impossible to directly get from experiments all interaction coefficients, but when it is only possible to observe the phenomenology of fixed or cyclic configurations. This corresponds in genetics to the phenotypic observation of stationary expression behaviors without experimental measure of the inhibitory and activatory coefficients of promoters and repressors. The possibility to obtain (even in an equivalence class) a sketch of the interaction matrix permits to construct (by randomizing the unknown coefficients of M) more complicated interaction matrices, then to test if they still have the observed states as fixed configurations, finally keep or reject these tested matrices and propose further experimental strategies using bio-arrays for refining the knowledge about the genetic network interaction structure.

7. REFERENCES

[1] J. Mattes, M. Richard, and J. Demongeot, "Tree representation for image matching and object recognition," *Lecture Notes in Comp. Sc.*, 1568, pp. 298-309, 1999.

[2] J. Demongeot, F. Estève, and P. Pachot, "Comportement asymptotique des systèmes : applications en biologie," *Rev. Int. Syst.*, 2, pp. 417-438, 1988.

[3] L. Mendoza, and E.R. Alvarez-Buylla, "Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis," *J. Theoret. Biology*, 193, pp. 307-319, 1998.

[4] M. Delbrück, "Unités biologiques douées de continuité génétique," *Colloques CNRS*, Paris, 8, pp. 33-35, 1949.

[5] J. Demongeot, M. Kaufmann, and R. Thomas, "Positive feedback circuit and memory," *C.R.A.S.*, 323, pp. 69-79, 2000.

[6] R. Thomas, "On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations," *Synergetics*, 9, pp. 1-23, 1980.

[7] S. Kauffman, "The Origins of Order," *Oxford University Press*, Oxford, England, 1993.

[8] D. Thieffry, M. Colet, and R. Thomas, "Formalization of regulatory networks : a logical method and its automatization," *Math. Mod. Sc. Comp.*, 2, pp. 144-151, 1993.

[9] E. Plahte, T. Mestl, and S.W. Omholt, “Feedback loops, stability and multi-stationarity in dynamical system,” *J. Biol. Syst.*, 3, pp. 409-414, 1995.
[10] J. Demongeot, “Multi-stationarity and cell differentiation,” *J. Biol. Syst.*, 6, pp. 1-2, 1998.
[11] E.H. Snoussi, “Necessary condition for multi-stationarity,” *J. Biol. Syst.*, 6, pp. 3-10, 1998.
[12] J.L. Gouzé, “Positive and negative circuits in dynamical systems,” *J. Biol. Syst.*, 6, pp. 11-16, 1998.
[13] J. Demongeot, J. Aracena, S. Ben Lamine, S. Meignen, A. Tonnelier, and R. Thomas, “Dynamical systems and biological regulations,” *Complex systems*, E Goles & S Martinez eds., Kluwer, Amsterdam, pp. 107-151, 2001.
[14] J. Aracena, S. Ben Lamine, M.A. Mermet, O. Cohen, and J. Demongeot, “Mathematical modelling in genetic networks, *BIBE 2000*, N. Bourbakis ed., IEEE Proc., pp. 141-149, 2000.
[15] J. Demongeot, J. Aracena, F. Thuderoz, T.P. Baum, and O. Cohen, “Genetic regulation network”, C. R. Acad. Sc., in press.
[16] O. Cinquin, and J. Demongeot, “Positive and negative feedback : striking a balance between necessary antagonists,” *J. Theoret. Biol.*, 215, in press.

Annex 2

Genetic regulation networks: circuits, regulons and attractors. Demongeot J, Aracena J, Thuderoz F, Baum TP, Cohen O. C. R. Acad. Sci. Biologies. 2003. 326:171-188.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 326 (2003) 171–188
COMPTES RENDUS BIOLOGIES

Biological modelling / Biomodélisation

Genetic regulation networks: circuits, regulons and attractors

Réseaux de régulation génétique : circuits, régulons, attracteurs

Jacques Demongeot *, Julio Aracena, Florence Thuderoz, Thierry-Pascal Baum, Olivier Cohen

Institut universitaire de France & CNRS TIMC–IMAG, Faculty of Medicine, 38700 La Tronche, France

Received 15 July 2002; accepted 4 December 2002

Presented by Jacques Ricard

Abstract

We deal in this paper with the concept of genetic regulation network. The genes expression observed through the bio-array imaging allows the geneticist to obtain the intergenic interaction matrix **W** of the network. The interaction graph G associated to **W** presents in general interesting features like connected components, gardens of Eden, positive and negative circuits (or loops), and minimal components having 1 positive and 1 negative loop called regulons. Depending on parameters values like the connectivity coefficient $K(\mathbf{W})$ and the mean inhibition weight $I(\mathbf{W})$, the genetic regulation network can present several dynamical behaviours (fixed configuration, limit cycle of configurations) called attractors, when the observation time increases. We give some examples of such genetic regulation networks and analyse their dynamical properties and their biological consequences. ***To cite this article: J. Demongeot et al., C. R. Biologies 326 (2003).***

Résumé

Cet article porte sur le concept de réseau de régulation génétique. L'expression génique, fournie par l'imagerie *bio-array*, permet d'obtenir la matrice **W** d'interaction intergénique du réseau. Le graphe d'interaction G associé à **W** présente en général des caractéristiques importantes telles que composantes connexes, jardins d'Eden, circuits (ou boucles) positifs et négatifs, ainsi que composants minimaux possédant une boucle négative et une boucle positive, appelés régulons. En fonction des valeurs de certains paramètres, tels que le coefficient de connectivité $K(\mathbf{W})$ et le poids moyen d'inhibition $I(\mathbf{W})$, le réseau de régulation génétique peut présenter différents comportements dynamiques (configuration fixe ou cycle limite de configurations) appelés attracteurs, lorsque le temps d'observation augmente. Nous donnerons des exemples de tels réseaux et analyserons leurs propriétés dynamiques, ainsi que leurs conséquences biologiques. Dans la partie consacrée à l'acquisition d'images bio-array, nous rappelons rapidement quelles sont leurs caractéristiques en termes de bruit et de signal et nous proposons une méthode (dite de l'emboutissage gaussien) permettant de les standardiser. Ensuite, nous donnons une méthode (dite des corrélations directionnelles) permettant d'extraire, à partir des images de co-expression des gènes, la matrice d'interaction inter-génique **W** liée à l'activité du réseau étudié. Puis, après description des caractéristiques majeures de son graphe associé G, nous donnons

* Corresponding author.
E-mail address: Jacques.Demongeot@imag.fr (J. Demongeot).

1631-0691/03/$ – see front matter
doi:10.1016/S1631-0691(03)00069-6

une suite de propositions, lemmes et théorèmes permettant de faire le lien entre la phénotypie observée (configurations fixes ou cycles de configurations au cours du cycle cellulaire des tissus étudiés) et les contraintes qu'elle exerce sur la structure interne de **W** et donc de G. Les deux résultats majeurs sont que l'existence d'au moins une boucle positive est une condition nécessaire de l'existence de plus d'un attracteur et que, si le nombre N de gènes est suffisamment grand et que $K(\mathbf{W}) = 2$, alors le nombre total $A(\mathbf{W})$ d'attracteurs est de l'ordre de $\sqrt{N}$ et de toute manière inférieur à 2^m, où m est le nombre de boucles positives du réseau de régulation génique. Nous donnons ensuite les exemples des réseaux de régulation contrôlant la floraison d'*Arabidopsis thaliana*, la gastrulation, la fonction lytique du phage μ et la préséance des bourgeons axillaires de *Bidens pilosa*, dans lesquels nous retrouvons la mise en œuvre des principales notions introduites dans les parties précédentes de l'article. ***Pour citer cet article : J. Demongeot et al., C. R. Biologies 326 (2003).***

Keywords: genetic regulation network; intergenic interaction matrix; positive loops; regulons; attractors

Mots-clés : réseau de régulation génétique ; matrice d'interaction intergénique ; boucles positives ; régulons ; attracteurs

1. Introduction

During the recent years, the rapid development of the bio-arrays techniques [1] based on isotopic or fluorescent activity of hybridised DNA chips allowed the biologist to give to a grey level peak the signification of an expression rate for the genes studied in the bio-array. If we repeat this acquisition at different times of the cell cycle for different cells of a same tissue, we can calculate correlations between the genes expression rate and hence we are able to make explicit a matrix **W** called the intergenic interaction matrix, representing the repression and induction influences a gene can exert on other genes.

1.1. The raw data from the bio-array imaging

The first encountered problem with the bio-array image is the noise and we have to low-pass it in order to suppress the high-frequency noise (see Fig. 1). The result of this pre-treatment is a better separation of

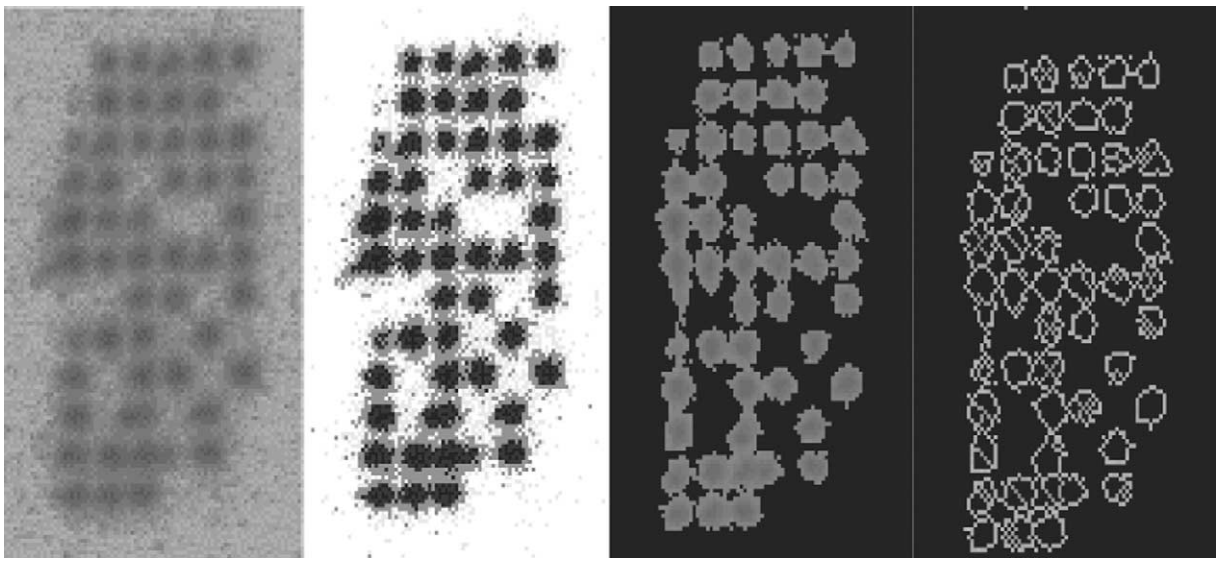


Fig. 1. (a) Raw data; (b) low-pass filtering; (c) watershed segmenting; (d) contouring.

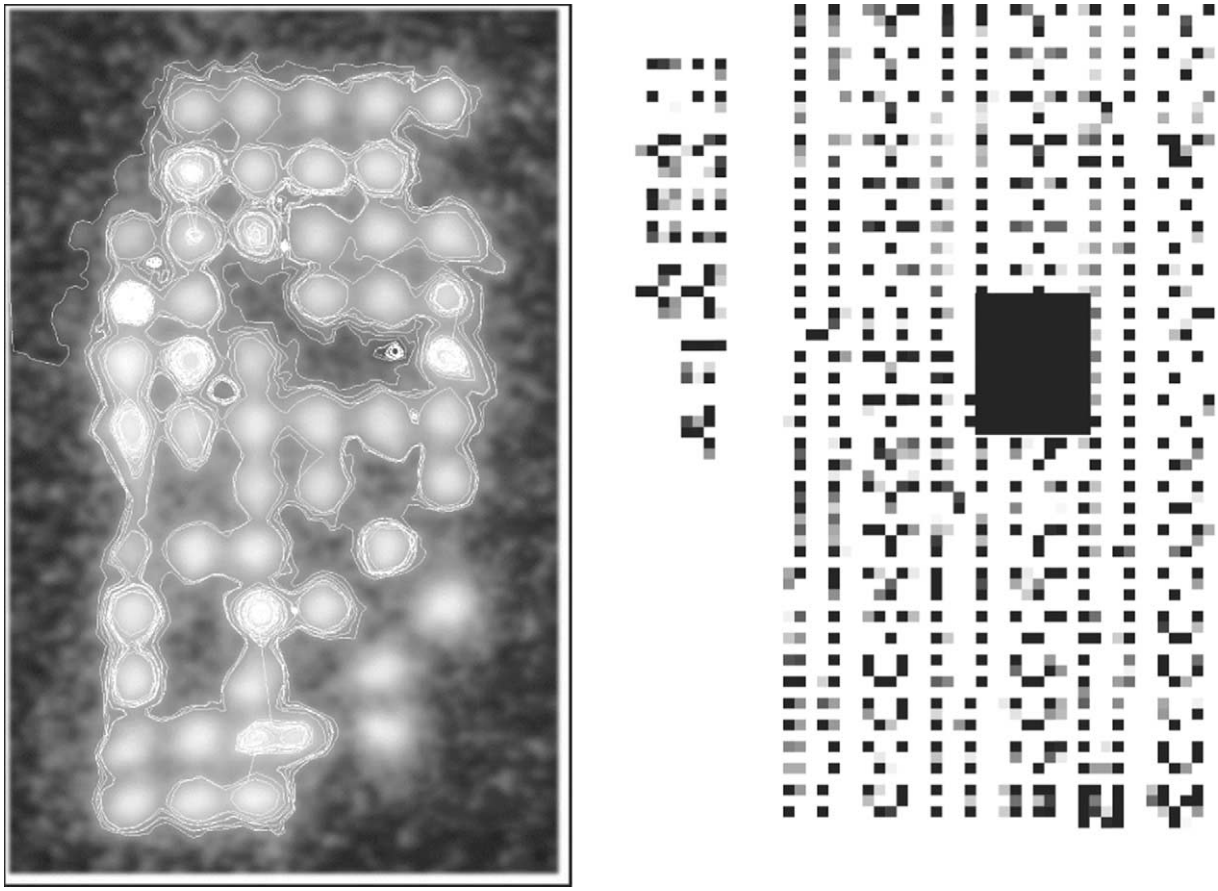


Fig. 2. 'Gaussian-stamping' contours (left) and standardized bio-array image (black rectangle right corresponds to the left image).

the isotopic activity peaks, allowing a watershed separation and contouring [2,3]. We can also apply a more accurate segmentation and contouring method called potential-Hamiltonian or 'Gaussian-stamping' method. Let us remark that the peaks are about Gaussian, with a weak kurtosis and skewness, allowing in particular the respect of the conservation 'law': 2/3 of the activity peak are concentrated into the set of points (x, y) where the Gaussian curvature $C(x, y)$ vanishes, i.e. inside the maximum gradient line, called the characteristic line of the peak. Then it is possible to neglect the part of the peak outside the projection of this line, whose equation is $C(x, y) = \partial^2 g/\partial x^2 \partial^2 g/\partial y^2 - (\partial^2 g/\partial x \partial y)^2 = 0$, $g(x, y)$ denoting the grey level function at the pixel of coordinates (x, y).

We are thus led to consider the new height function $C(x, y)$ instead of the function $g(x, y)$ and its level line $C(x, y) = 0$. A new algorithm has been proposed in [1] to obtain automatically the characteristic line and, after, by integrating inside the projection of this line on the (x, y) plane and multiplying by 3/2 the obtained result, to standardize the estimated activity in terms of a bio-image with small squares symbolizing in grey levels the degree of hybridisation of the cDNA's expressing the regulation of a glioma tissue (Fig. 2). From such bio-array images acquired in cells of the same tissue at different times of the cell cycle, we can study the interactions between genes by estimating an interaction matrix.

2. Some rigorous results about the network attractors

2.1. The interaction matrix **W**

The interaction matrix **W** is similar to the synaptic weight matrix, which rules the relationships between neurons in a neural network. The general coefficient w_{ik} of such an interaction matrix **W** is equal to $+1$ (resp -1, 0) if the gene G_k activates (resp inhibits, does not influence) the gene G_i, the state x_i of the gene G_i being equal to $+1$ (resp -1), if it is (resp is not) expressed. In the case of small regulatory genetic systems (called *operons*), the knowledge of such a matrix **W** permits to make explicit all possible stationary behaviours of the organisms having the corresponding genome. The change of state of gene G_i between t and $t+1$ obeys a threshold rule: $x_i(t+1) = H(\sum_{k=1,n} w_{ik}x_k(t) - b_i)$ or $x(t+1) = H(\mathbf{W}\mathbf{x}(t) - b)$, where H is the sign step function ($H(y) = 1$, if $y \geqslant 0$ and $H(y) = -1$, if $y < 0$) and the b_is are threshold values. When t is increasing, the genes states reach a stable set of configurations (a fixed configuration or a cycle of configurations), called attractor of the genetic network dynamics. For example, in the regulatory network that regulates the *Arabidopsis thaliana* flower morphogenesis, the interaction matrix is a $(11, 11)$-matrix with only 22 non-zero coefficients (see Fig. 3). This matrix presents $P(\mathbf{W}) = 4$ positive loops and $A(\mathbf{W}) = 4$ attractors (see Section 3.1). Hence it is in general of a great biological interest and relevance to determine matrices having characteristic properties like (i) a minimal number of non-zero coefficients for a given set of attractors (fixed points or cycles) or (ii) a minimal number $P(\mathbf{W})$ of positive loops, controlling the number $A(\mathbf{W})$ of attractors and their stability (cf. [4–9] for the continuous case [10–12] and for the discrete one). In the following, we intend to partly solve the problems taken above by giving necessary and sufficient conditions to obtain properties (i) and (ii). In order to calculate the w_{ik}s, we can either determine the s-directional correlation $\rho_{ik}(s)$ between the state vector $\{x_k(t-s)\}_{t\in C}$ of gene j at time $t-s$ and the state vector $\{x_i(t)\}_{t\in C}$ of gene i at time t, t varying during the cell cycle C of length $M = |C|$ and corresponding to observation times of the bio-array images:

$$\rho_{ik}(s) = \left(\sum_{t\in C} x_k(t-s)x_i(t) - \sum_{t\in C} x_k(t-s)\sum_{t\in C} x_i(t)/M\right) \Big/ \sigma_k(s)\sigma_i(0)$$

where

$$\sigma_k(s) = \left(\sum_{t\in C} x_k(t-s)^2 - \left(\sum_{t\in C} x_k(t-s)\right)^2 \Big/ M\right)^{1/2}$$

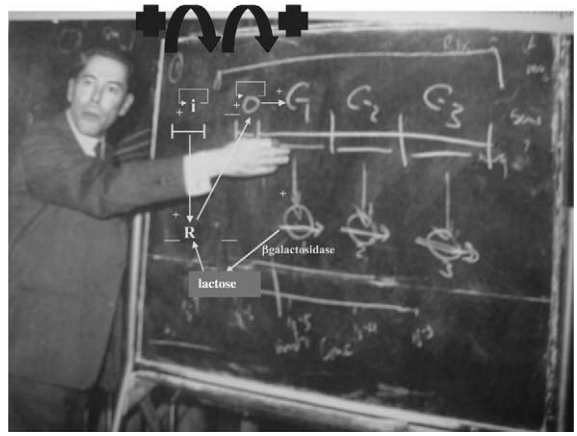


Fig. 3. The lactose operon exhibited by Jacques Monod.

and then take $w_{ik} = \text{sign}(\sum_{s=1,\ldots,m} \rho_{ik}(s)/M)$, if $|w_{ik}| > \eta$, and $w_{ik} = 0$, if $|w_{ik}| < \eta$, where η is a de-correlation threshold, or identify the system with a Boolean neural network. When it is impossible to so obtain all the coefficients of $\mathbf{W}$ (neither from the literature nor from such calculations), it is possible to complete $\mathbf{W}$: we choose randomly the missing coefficients by respecting the connectivity coefficient $K(\mathbf{W}) = I/N$, the ratio between the number I of interactions and the number N of genes, and the mean inhibition weight $I(\mathbf{W}) = R/I$, the ratio between the number of inhibitions (or repressions) R and $I \cdot K(\mathbf{W})$ is in general between 1.5 and 3 and $I(\mathbf{W})$ between $1/3$ and $2/3$ for many known operons or regulatory networks (lactose operon, Cro operon for the phage δ, lysogenic/lytic operon for the phage μ, gastrulation and *Arabidopsis thaliana* flowering regulatory networks...).

If G is the interaction graph associated to $\mathbf{W}$, then we call connected component C of G each set of genes C such that there is a path between every pair of genes of C along a sequence of arcs of G. A Garden of Eden is a gene receiving no arc, but influencing at least one other gene. A regulon is a connected component of G having exactly one positive (auto-catalysis) and one negative loop, these loops sharing the auto-catalysed node. In the lactose operon (see its G in Fig. 3), $K(\mathbf{W}) = 8/6$, $I(\mathbf{W}) = 3/8$, $P(\mathbf{W}) = 2$, $A(\mathbf{W}) = 2$ (βgal-activated and inactivated states), and G has one connected component and one regulon.

2.2. Some definitions and notations

In the following, we give some definitions about the rigorous mathematical description of the discrete Boolean networks used to describe the genes interaction dynamics, and their associated graph G and matrix $\mathbf{W}$. Then we will present some theoretical results with rigorous proofs only for the first and last results in order to show what kind of reasoning we have to perform and we will refer to [13–16] for complete demonstrations. Let us consider a graph G = (V, E), where $V = \{1, \ldots, n\}$ is the set of nodes and E is the set of arcs. Let $\mathbf{W} = (w_{ik})$ be a real (n, n)-matrix. We call G the incidence graph of $\mathbf{W}$, if (k, i), the arc going from k to i, belongs to E, then $w_{ik} \neq 0$. By extension, $\mathbf{W}$ will also be called the incidence matrix of G. We will define the sign of an arc (k, i), denoted by $\text{sign}((k, i))$, as the sign of w_{ik}. Let us denote by $\Gamma^-(i)$ (resp $\Gamma^+(i)$) the set of nodes $\{i_1, i_2, \ldots, i_{k(i)}\}$ such that (i_j, i) (resp (i, i_j)) belongs to E, for each $j = 1, \ldots, k(i)$. We will say that a set of arcs $C = \{e_1, e_2, \ldots, e_r\}$ is a chain if each e_k in C has a node belonging to e_{k-1} and the other one belonging to e_{k+1}. We will say that C is a simple (resp elementary) chain if the arcs (resp nodes) are different. In the sequel we will understand by chain a simple and elementary chain. In the same way we will call C a path if $e_k = (i_k, i_{k+1})$ implies $e_{k+1} = (i_{k+1}, i_{k+2})$, for all $k = 1, \ldots, r$, that is to say the final node of each arc is the beginning node of the next arc in C. The sign of a path or a chain C (denoted by sign(C)) is positive if the number of negative arcs of C is even and negative if this number is odd. A cycle (resp circuit or loop) is defined as a chain (resp path) where each of the two extremities of any arc belongs to two and only two arcs. For simplicity of notation, we will say that a node i belongs to a cycle C if there exists a node j such that (i, j) or (j, i) belongs to C. Every other definition of graph theory used here will be consistent with that in [17,18]. We will call a circuit or cycle C negative (resp positive) if sign(C) is negative (resp positive). Define now a discrete state regulatory network, acting on the set of states $\{-1, 1\}$, here and subsequently denoted by N, as the 4-uple $N = (G, \mathbf{W}, b, \text{sign})$, where G is the incidence graph of $\mathbf{W}$, b is a threshold real vector and the local transition function is given by $x_i(t+1) = \text{sign}(\sum_{k=1,\ldots,n} w_{ik} x_k(t) - b_i)$, $\forall \mathbf{x} \in \{-1, 1\}^n$, where $\text{sign}(u) = 1$ if $u \geqslant 0$ and $\text{sign}(u) = -1$ otherwise. The sequential iteration consists to update the nodes one by one in a prescribed periodic update $I = (i(1), i(2), \ldots, i(n))$, where $\{i(1), i(2), \ldots, i(n)\} = \{1, 2, \ldots, n\}$, that is to say, starting with a given $\mathbf{x}(0) = (x_1(0), \ldots, x_n(0))$ in $\{-1, 1\}^n$, we generate a sequence of iterates:

$$x_{i(k)}(t) = \text{sign}\left(\sum_{j<k} w_{i(k)i(j)} x_{i(j)}(t) - b_{i(k)} + \sum_{j \geqslant k} w_{i(k)i(j)} x_{i(j)}(t-1) - b_{i(k)} \right), \quad \forall k \in \{1, \ldots, n\}$$

Now, the parallel iteration consists in updating all the nodes synchronously:

$$x_i(t+1) = \text{sign}\left(\sum_{k=1,\ldots,n} w_{ik} x_k(t) - b_i \right), \quad \forall i \in \{1, \ldots, n\}, \quad \text{with} \quad \mathbf{x}(0) \in \{-1, 1\}^n$$

We shall say that **x** is a fixed point if it is invariant under the application of the complete sequence of updates. Observe that the kind of iteration does not change the set of fixed points, but only change their attraction basins. In the following we will use systematically the parallel iteration.

2.3. Relations between positive and negative cycles, and fixed points

In the sequel, we will assume that the graph G is connected, since otherwise one can apply the results to each of connected components of G. In addition, we will suppose with no loss of generality that $|\Gamma^-(i)| > 0$, for all $i \in V$. Since otherwise, if there exists a node $i \in V$ such that $\Gamma^-(i)$ is empty, then we can assume that the arc (i, i) exists in E; in this way the dynamics of both networks are the same. It evidently follows from this property that there exists at least one circuit C in G (possibly a circuit of the form (i, i)). Finally, we suppose that the graph G and the matrix **W** have a quasi-minimal structure, that is to say, all (j, i), such as $i \neq j$, belong to E (or equivalently $w_{ij} \neq 0$, if $i \neq j$), if there exists $\mathbf{x} \in \{-1, 1\}^n$, such that:

$$\operatorname{sign}\left(\sum_k w_{ik}x_k - b_i\right) \neq \operatorname{sign}\left(\sum_{k\neq j} w_{ik}x_k - b_i\right)$$

Hence, we have the following necessary condition to have a quasi-minimal structure:

$$-\sum_k |w_{ik}| < b_i \leqslant \sum_k |w_{ik}|, \quad \forall i \in 1, \ldots, n$$

The following property will be very useful in the following for characterizing a cycle.

Proposition 1. *A cycle* C *is positive if and only if there exists a vector* $\mathbf{x} \in \{-1, 1\}^n$ *such that for all* $(k, i) \in C$, $\operatorname{sign}(w_{ik}) = x_i x_k$ *or equivalently, for all* $(k, i) \in C$, $x_i = \operatorname{sign}(w_{ik})x_k$ (1).

Proof. Let C be a positive cycle and $i(0)$ a fixed node belonging to C. Let us enumerate the nodes belonging to C by $i(0), i(1), \ldots, i(j)$, such that $\forall k = 0, \ldots, j$, $(i(k-1), i(k))$ or $(i(k), i(k-1)) \in C$ (by identifying j and -1). Finally, let us define the vector **x** as follows:

$$x_{i(0)} = 1 \quad \text{and} \quad x_{i(k)} = \operatorname{sign}(w_{i(k)i(k-1)})x_{i(k-1)} \quad \text{if } \big(i(k-1), i(k)\big) \in C \quad \text{or}$$
$$x_{i(k)} = \operatorname{sign}(w_{i(k-1)i(k)})x_{i(k-1)} \quad \text{if } \big(i(k), i(k-1)\big) \in C, \ \forall k = 1, \ldots, j$$

Obviously **x** is satisfying equation (1). Hence –**x** satisfies equation (1) too. Finally, it is direct that there does not exist another vector $\mathbf{y} \notin \{\mathbf{x}, -\mathbf{x}\}$ that satisfies equation (1).

Let C be now a negative cycle, and let us suppose that equation (1) is true, then $\prod_{(j,i)\in C} \operatorname{sign}(w_{ij}) = \prod_j x_j \prod_i x_i = (\prod_j x_j)^2$, but $\operatorname{sign}(C) = \prod_{(j,i)\in C} \operatorname{sign}(w_{ij}) < 0$, which is contradictory. □

Theorem 1. *Given* N, *if all cycles of the incidence graph* G *are positive, then there exists a vector* $\mathbf{x} = (x_1, \ldots, x_n) \in \{-1, 1\}^n$ *such that* **x** *and* $-\mathbf{x} = (-x_1, \ldots, -x_n)$ *are fixed points of* N.

Remark. There are two remarkable fixed points having by construction a non-frustration property, that is on each cycle the sign changes of x_is are identical to the sign changes of the arcs. For other possible fixed points, there is at least one cycle for which sign changes are frustrated.

Theorem 2. *If all circuits of the incidence graph* G *are negative, then* N *has no fixed points.*

2.4. Minimal regulatory networks

The previous results allow us to characterize some minimal regulatory networks. The following propositions constitute examples of minimal regulatory networks. They solve in part the inverse problem consisting in the description of **W** only from the knowledge of a phenotypic **x** observed from bio-array images.

Proposition 2. *Let* N *having n nodes and n connections, a necessary and sufficient condition of existence of a fixed point* **x** *is the existence of a positive circuit. In this case,* **x** *and* $-\mathbf{x}$ *are both fixed points. Hence we can characterize the set of minimal* Ns *having* **x** *as fixed point.*

Proposition 3. *Given a state vector* **x**, *the set of minimal networks* $\mathrm{N} = (\mathrm{G}, \mathbf{W}, b, \mathrm{sign})$ *having* **x** *as fixed point is given by the following conditions*:

(1) $w_{ik} = \alpha_{ik} x_i x_k$, *where* $\alpha_{ik} \geqslant 0$ *and, for all* i, *there exists* $k(i)$ *such that* $\alpha_{ik(i)} \neq 0$ *and*
(2) $-|\alpha_{ik(i)}| < b_i \leqslant |\alpha_{ik(i)}|$.

Proposition 4. *Let* N *with n nodes and* $n+1$ *connections, a necessary and sufficient condition for existence of an attractor of all points parallel iterated, is a negative circuit and a positive circuit intersecting.*

2.5. Fixed points bounds in regulatory networks

Given N, let C be a positive circuit of N, then by Proposition 1, there exists $\mathbf{x} \in \{-1, 1\}^{|\mathrm{V(C)}|}$, such that **x** and $-\mathbf{x}$ satisfy the equation: $\forall (k, i) \in \mathrm{C}$, $\mathrm{sign}(w_{ik}) = x_i x_k$. We denote $\mathbf{u}(\mathrm{C}) \in \{-1, 1\}^{|\mathrm{V(C)}|}$ the vector defined by: $\mathbf{u}(\mathrm{C}) = \mathbf{x}$ (resp $-\mathbf{x}$), if $x_{i(0)} = 1$ (resp -1), where $i(0) = \min\{i \mid i \in \mathrm{C}\}$.

Lemma 1. *Given* N *and* **y** *a fixed vector of* N, *then for all* $i \in \mathrm{V}$, *there exists a positive circuit* $C(i)$ *in* G *such that for all* k *in* $C(i)$, $y_k = u(C(i))_k$ *or for all* k *in* $C(i)$, $y_k = -u(C(i))_k$.

Theorem 3. *If* m *is the total number of positive circuits of* N, *then the number of fixed points of* N *is* $\leqslant 2^m$, *and this upper bound is reached if and only if for all circuits* C *of* N *there does not exist an arc* (k, i) *in* C^{C} *ending in* C (*there is no garden of Eden* k *pending to* C).

Remark. We have to notice that the condition concerns the number of circuits and not of cycles, these last being in general very more numerous.

2.6. Asymptotic mean value for the number of fixed configurations (fixed vectors or limit cycles of vectors) in the case $K(\mathbf{W}) = 2$

Let us consider now a network N having n nodes and $2n$ connections such as $K(\mathbf{W}) = 2$. We search for a mean value of the number of fixed configurations, when n is growing to infinity.

Lemma 2. *For any graph* G *having m non-oriented edges, the mean number of oriented edges we can define on* G *from the non-oriented configuration is equal to* $4m/3$.

Proof. Let us note $\langle o \rangle$ the mean number of oriented edges we can construct from a configuration of m non-oriented edges; then, if exactly k from the m non-oriented edges are decomposed into two oriented opposite connections, we have $\mathrm{C}_m^{m-k} 2^{m-k}$ different ways to dispatch the not double connections into the $(m-k)$ other non-oriented edges; hence we can write: $\langle o \rangle = \sum_{k=0}^{m} (k+m)\, \mathrm{C}_m^{m-k} 2^{m-k} / \sum_{k=0}^{m} \mathrm{C}_m^{m-k} 2^{m-k} = 4m/3$. □

Theorem 4. *If the network* N *has n nodes and* $K(\mathbf{W})n$ *connections, with* $K(\mathbf{W}) = 2$, *then the expectation of the number of fixed configurations of* N *is* $n^{1/2}$, *if n is sufficiently large.*

Proof. Following [19], if the connections of N are random, and if the mean number c of non-oriented edges per node is equal to $3/2$, then the random variables X_i equal to the number of disjoint cycles of length i of N are independent and Poissonnian with parameter $\lambda(i) = 2^{i-1}/i$, if n is sufficiently large. From Lemma 2, we are just in this case, because we have $2n = 4m/3$ connections, hence $m = 3n/2$ and $c = m/n = 3/2$. Then we have, for the mean number $\langle f\rangle$ of fixed configurations of N: $\langle f\rangle = \sum_{s=0}^{n}\sum_{k=s}^{n}\sum_{\sigma\in\Omega(s,k)} A(\sigma)\Pi_\sigma$, where $\Omega(s,k) = \{\sigma = (s(1),\ldots,s(n))/s(i) \geqslant 0, \sum_{i=1}^{n} s(i) = s, \sum_{i=1}^{n} is(i) = k\}$, $\Pi_\sigma = P(\{X_i = s(i), s(i) \geqslant 0, \sum_{i=1}^{n} s(i) = s, \sum_{i=1}^{n} is(i) = k\}) = \mathrm{e}^{-\sum\lambda(i)}\prod_{i=1}^{n}\lambda(i)^{s(i)}/s(i)!$ is the probability to have the X_is equal each to $s(i)$, and $A(\sigma)$ is the mean number of fixed configurations, when each X_is is equal to $s(i)$.

We will now evaluate the expectation $A(\sigma)$. Each disjoint positive circuit bringing two fixed points (Theorem 3 above), an isolated positive non-circuit cycle bringing also two fixed points and an isolated negative circuit bringing one limit cycle, we can first calculate $A(0,\sigma)$, the expected number of fixed configurations in the case where we have only disjoint positive circuits, the rest of the nodes depending on these circuits (and hence their states being fixed by the states of the circuit): $A(0,\sigma) = B(0,\sigma)/D(\sigma)$, where $B(0,\sigma) = 2^s$ (number of fixed points of $s = \sum_{i=1}^{n} s(i)$ disjoint positive circuits, from Theorem 3 above) $\times\ 2^k$ (number of different signs for each of the $k = \sum_{i=1}^{n} is(i) \leqslant n$ nodes involved in the s circuits) $\times$ [2^s (number of different directions – left or right – for each of the s circuits)$/2^s$ (reduction factor for having only positive circuits)] $\times\ N(\sigma)$, $D(\sigma) = 2^k$ (number of different directions for each of the k connections) $\times\ 2^k$ (number of different signs for each of the k connections) $\times\ N(\sigma)$, where $N(\sigma)$ is the number of choices for the s disjoint cycles: $N(\sigma) = \mathrm{C}_k^{s(1)1,\ldots,s(n)n}\prod_{i=1}^{n}(i-1)!^{s(i)}/2^s$. $N(\sigma)$ just equals the number of choices of k nodes in $s(1)$ subsets of size $1,\ldots,s(n)$ of size n times the number of choices of different loops (without multiple points) connecting the vertices inside each of these subsets.

In the same way, we can calculate $A(1,\sigma)$ (resp $A(j,\sigma)$) the expected number of attractors of N in the case where we have among the s disjoint cycles 1 (resp j) isolated positive non-circuit cycles (bringing two fixed vectors) or isolated negative circuits (bringing 1 fixed cycle).

We have also:

$$A(1,\sigma) = B(1,\sigma)/D(\sigma)$$

where

$$\begin{aligned}B(1,\sigma) = {} & 2^{s-1}(2^{s-1}/2^{s-1})N(\sigma)\big[2^{k-1}s(1)(2^1 2^1 2^1 + 2^1 2^1 - 2^1 2^1 2^1)/2^1 + \cdots \\ & + 2^{k-i}s(i)(2^1 2^i 2^i + 2^i 2^1 - 2^1 2^i 2^1)/2^1 + \cdots + 2^{k-n}s(n)(2^1 2^n 2^n + 2^n 2^1 - 2^1 2^n 2^1)/2^1\big]\end{aligned}$$

where $s(i)2^1 2^i 2^i/2^1$ is just the number (2^1) of fixed points of 1 positive cycle (circuit or not) of length i times the number of such configurations $s(i)(2^i 2^i/2^1)$, $s(i)2^i 2^1/2^1$ is the number (1) of attractors of 1 isolated negative circuit of length i times the number of such configurations $s(i)$ $(2^i 2^1/2^1)$ and $-s(i)2^1 2^i 2^1/2^1$ is the number (2^1) of fixed points of 1 positive circuit (already counted in $B(0,\sigma)$) times the number of such configurations $s(i)$ $(2^i 2^1/2^1)$; 2^{s-1} $(2^{s-1}/2^{s-1})N(\sigma)2^{k-i}$ is equal to the number of configurations of $s-1$ positive circuits with $s(1)$ of length $1,\ldots,s(i)-1$ of length $i,\ldots,s(n)$ of length n. Then we have: $A(2,\sigma) = B(2,\sigma)/D(\sigma)$, where

$$\begin{aligned}B(2,\sigma) = {} & 2^{s-2}(2^{s-2}/2^{s-2})N(\sigma)\big[2^{k-2}s(1)^2\big(2^2 2^2 2^2 - 2(2^2 2^2 2^1 - 2^1 2^2 2^1) + 2^2 2^1\big)/2^2 + \cdots \\ & + 2^{k-i-j}s(i)s(j)\big(2^2 2^{i+j} 2^{i+j} - (2^2 2^{i+j} 2^i 2^1 - 2^1 2^{i+j} 2^i 2^1) - (2^2 2^{i+j} 2^j 2^1 - 2^1 2^{i+j} 2^j 2^1) \\ & + 2^{i+j} 2^2\big)/2^2 + \cdots + 2^{k-2n}s(n)^2\big(2^2 2^{2n} 2^{2n} - 2(2^2 2^{2n} 2^n 2^2 - 2^1 2^{2n} 2^n 2^2) + 2^{2n} 2^2\big)/2^2\big]\end{aligned}$$

where $s(i)s(j)(2^2 2^{i+j} 2^{i+j} - (2^2 2^{i+j} 2^i 2^2 - 2^1 2^{i+j} 2^i 2^2) - (2^2 2^{i+j} 2^j 2^2 - 2^1 2^{i+j} 2^j 2^2) + 2^{i+j} 2^2)/2^2$ is just the number of the fixed configurations of a couple made of positive not circuit cycles or negative circuits, the $(s-2)$ remaining cycles being positive circuits, by paying attention to the fact that the fixed configurations of the couple

of a positive not circuit cycle combined with a positive circuit $(s(i)s(j)2^2 2^{i+j}(2^i + 2^j)2^2)$ have been already counted both in $B(1,\sigma)$ and in $B(0,\sigma)$ and hence have to be taken away (by using sign $-$) from the sum $s(i)s(j)(2^2 2^{i+j} 2^{i+j} + 2^1 2^{i+j} 2^i 2^1 + 2^1 2^{i+j} 2^j 2^1 + 2^{i+j} 2^2)/2^2$.

Finally, more generally, we have $A(j,\sigma) = B(j,\sigma)/D(\sigma)$, where

$$
\begin{aligned}
B(j,\sigma) &= 2^{s-j}(2^{s-j}/2^{s-j})N(\sigma) \\
&\quad \times \Bigg[\sum_{\zeta \in I} 2^{k-r(\zeta)} \Bigg[2^j 2^{r(\zeta)} 2^{r(\zeta)} + \sum_{m=1}^{j} 2^{j-1} 2^{r(\zeta)-i(m)} 2^{r(\zeta)-i(m)} (2^{i(m)} 2^1 - 2^1 2^{i(m)} 2^1) + \cdots \\
&\qquad + \sum_{\xi=(m(1),\ldots,m(v)) \in \{1,\ldots,n\}} v 2^{j-v} 2^{r(\zeta)-r(\xi)} 2^{r(\zeta)-r(\xi)} \big(2^{r(\xi)} 2^v - v(2^2 2^{r(\xi)} 2^v - 2^1 2^{r(\xi)} 2^v) \\
&\qquad + v(v-1)(2^2 2^{r(\xi)} 2^v - 2 \cdot 2^3 2^{r(\xi)} 2^v + 2^4 2^{r(\xi)} 2^v)/2 + \cdots \\
&\qquad + (-1)^v 2^v 2^{r(\xi)} 2^v \big) + \cdots + (-1)^j 2^j 2^{r(\zeta)} \Bigg] \Big/ 2^j \Bigg] \\
&= 2^{s-j}(2^{s-j}/2^{s-j})N(\sigma) \sum_{I} \prod_{t=1}^{j} \big(2^{i(t)-1} - 1/2\big)^{u(i(t))}
\end{aligned}
$$

where $I = \{\zeta = (i(1),\ldots,i(j)) \in \{1,\ldots,n\}^j \mid \forall t = 1,\ldots,j$, the number $u(i(t))$ of cycles of size $i(t)$ satisfies: $0 < u(i(t)) \leqslant s(i(t))\}$, $j = \sum_{t=1}^{j} u(i(t))$, $r(\zeta) = \sum_{t=1}^{j} i(t)$, and $2^j 2^{r(\zeta)} 2^{r(\zeta)}/2^j$ is just the number (2^j) of fixed points of j positive cycles (circuits or not) of lengths $i(1),\ldots,i(j)$ multiplied by the number of such configurations of j positive cycles $(2^{r(\zeta)} 2^{i(1)+\cdots+i(j)}/2^j)$, $\sum_{m=1}^{j} 2^{j-1} 2^{r(\zeta)-i(m)} 2^{r(\zeta)-i(m)}$. $(2^{i(m)} 2^1 - 2^1 2^{i(m)} 2^1)/2^j$ being the number of attractors in a configuration where we have 1 negative circuit of length $i(m)$ among the $(j-1)$ other positive cycles (circuits or not) diminished by the number of the configurations having $(j-1)$ positive non-circuit cycles and $(k-j+1)$ positive circuits (already counted in $B(j-1,\sigma)$). The other terms of $B(j,\sigma)$ correspond to the number of fixed configurations of sub-graphs having $(k-j)$ positive circuits and j either positive non-circuit cycles or negative circuits, diminished by the number of already counted fixed configurations in the $B(m,\sigma)$s, for $m < j-1$, and not yet taken away. We remark to end the calculation of $B(j,\sigma)$ that:

$$
\begin{aligned}
&2^{k-r(\zeta)} \Bigg[\sum_{\xi=(m(1),\ldots,m(v)) \in \{1,\ldots,n\}} v 2^{j-v} 2^{r(\zeta)-r(\xi)} 2^{r(\zeta)-r(\xi)} \big(2^{r(\xi)} 2^v - v(2^2 2^{r(\xi)} 2^v - 2^1 2^{r(\xi)} 2^v) \\
&\qquad + v(v-1)(2^2 2^{r(\xi)} 2^v - 2 \cdot 2^3 2^{r(\xi)} 2^v + 2^4 2^{r(\xi)} 2^v)/2 + \cdots + (-1)^v 2^v 2^{r(\xi)} 2^v \big) \Bigg] \Big/ 2^j \\
&= \sum_{\xi=(m(1),\ldots,m(v)) \in \{1,\ldots,n\}} v 2^v 2^k 2^{r(\zeta)-r(\xi)} (1/2 - 1)^v = \sum_{\xi=(m(1),\ldots,m(v)) \in \{1,\ldots,n\}} v 2^k 2^{r(\zeta)-r(\xi)} (-1)^v
\end{aligned}
$$

By summing the $A(j,\sigma)$s and after the $A(\sigma)\Pi_\sigma$s, $\langle f \rangle$ is clearly of the order of $n^{1/2}$:

$$
\begin{aligned}
\langle f \rangle &= \sum_{s=0}^{n} \sum_{k=s}^{n} \sum_{\sigma \in \Omega(s,k)} \mathrm{e}^{-\sum \lambda(i)} \prod_{i=1}^{n} \lambda(i)^{s(i)}/s(i)! \sum_{j=0}^{n} A(j,\sigma) \\
&= \sum_{s=0}^{n} \sum_{k=s}^{n} \sum_{\sigma \in \Omega(s,k)} \big(\mathrm{e}^{-\sum \lambda(i)}/s! \big) \Bigg(s! \Big/ \prod_{i=1}^{n} s(i)! \Bigg) K_\sigma
\end{aligned}
$$

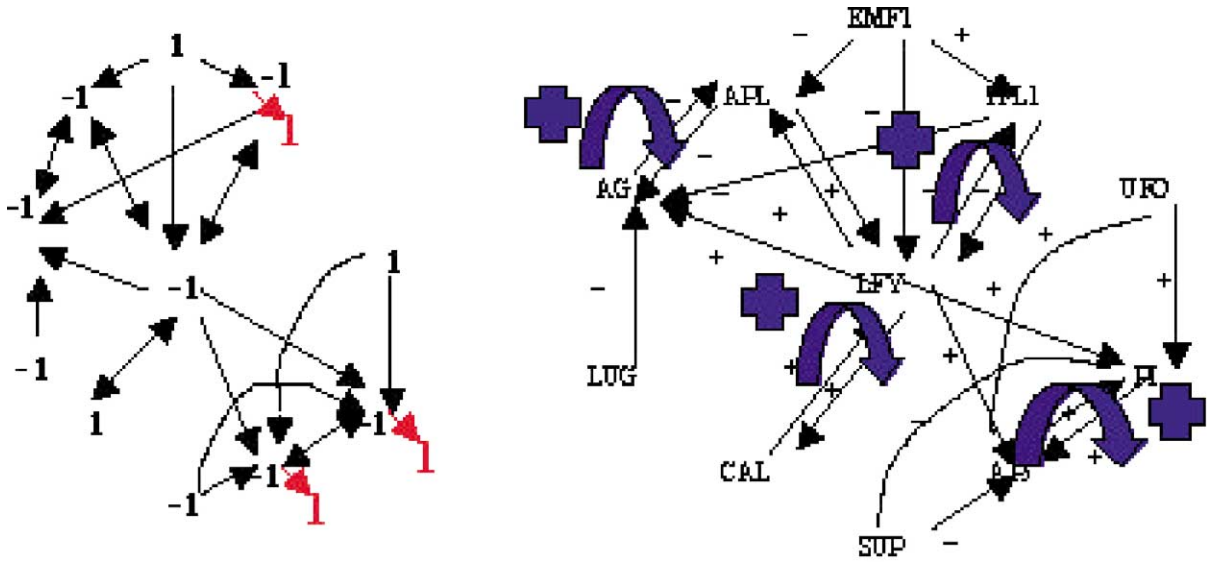


Fig. 4. Interaction graph of the flowering regulatory network of *Arabidopsis thaliana* (right) and an attractor of its Boolean dynamics (left).

where

$$K_\sigma = 2^{s-k}\prod_{i=1}^{n}\lambda(i)^{s(i)}(2^{i-1}-1/2+1)^{s(i)} = \prod_{i=1}^{n}(2^i/i)^{s(i)}(1+1/2^i)^{s(i)} \quad \text{and}$$

$$\sum_{i=1}^{n}\lambda(i) = \left(\sum_{i=1}^{n}2^i/i\right)/2 = \sum_{i=1}^{n}\int_0^2 x^{i-1}\,\mathrm{d}x/2 = \int_0^2\sum_{i=1}^{n}x^{i-1}\,\mathrm{d}x/2$$

$$= \int_0^2 (x^n-1)/(x-1)\,\mathrm{d}x/2 < 2^{n-1}$$

Then we have: $\langle f\rangle \sim \sum_{s=0}^{n} \mathrm{e}^{-\sum\lambda(i)}(\sum_{i=1}^{n}\lambda(i)+\ln n/2)^s/s! = \mathrm{O}(\sqrt{n})$. □

Remark. Theorem 4 corresponds to the Kaufmann's conjecture [20].

3. Examples of genetic regulation networks

3.1. The flowering regulatory network of Arabidopsis thaliana

If we consider the interaction graph of the flowering regulatory network of *Arabidopsis thaliana* (Fig. 4, right) [10], then we can easily define from it a Boolean dynamics, Fig. 4 (left) giving an example of attractor with final states (in bold red) different from the initial conditions. The characteristics of the associated interaction matrix **W** are: $K(\mathbf{W}) = 22/11 = 2$, $I(\mathbf{W}) = 10/22$, $P(\mathbf{W}) = 4$, $A(\mathbf{W}) = 4$ (corresponding to the 4 differentiated tissues of the flower, i.e. sepals, petals, stamens and carpels). **W** has two connected components and four gardens of Eden. $A(\mathbf{W})$ is well $\leqslant 2^4$ and in fact exactly equal to 2^2, where 2 is the number of connected components having at least one positive loop. Then we can recall that:

Fig. 5. The Watt regulator, the prototype of negative regulatory loop in cybernetics.

– in 1948, Delbrück [21] conjectured that positive loops in the interaction graph of a regulatory network was a necessary condition for cell differentiation, i.e. for the existence of multiple attractors of the genes expression; this conjecture has been written in a good mathematical context by Thomas in 1980 [22]; we have proved above that the positive loops were related to the observation of multiple attractors, which definitively gives to the positive loops another signification than to the negative ones, more related to the stability of the system (like in the classical Watt regulator, well known in cybernetics, cf. Fig. 5);
– in 1992, Kauffman [20] conjectured that the mean number of attractors for a Boolean genetic network with n genes and with connectivity 2 was of the order of $\sqrt{n}$ (see Theorem 4 above). This conjecture is now supported by real observations: we have about 35 000 genes in the human genome and about 200 different tissues, which can be considered as different attractors of the same dynamics. For *Arabidopsis thaliana*, there is $A(\mathbf{W}) = 4 \approx \sqrt{11}$ different tissues [10] and for the Cro operon of the phage λ, $K(\mathbf{W}) = 14/5 = 2.8$, $I(\mathbf{W}) = 9/14$, $A(\mathbf{W}) = 2 \approx \sqrt{5}$ (lytic and lysogenic attractors) [23,24], with Boolean [25] or discrete multi-level [26] models.

3.2. The gastrulation regulatory network

If we consider the regulatory network ruling the gastrulation in *Drosophila* (cf. Fig. 6 and [28]), it is easy to check that $K(\mathbf{W}) = 25/15$, $I(\mathbf{W}) = 5/15$, $P(\mathbf{W}) = 4$ and $A(\mathbf{W}) = 2$ (the corresponding cells being the ordinary ectoderm cell and the trapezoidal invagination cell called bottle cell). The regulation graph contains five connected components (among which three are singletons). In this case, the classical Kolmogorov–Rashevski–Turing models of reaction–diffusion [29–31] are well explaining the epigenetic part of the invagination at the start of the gastrulation, but only after the apparition of a new bottle cell presenting an apical constriction due to the change of intracellular balances ATP/ADP and GTP/GDP (whose ratios increase), due to the expression of the kinase (ADK and GDK) genes. This is due to a change of attractor basin by the bottle cell starting the gastrulation

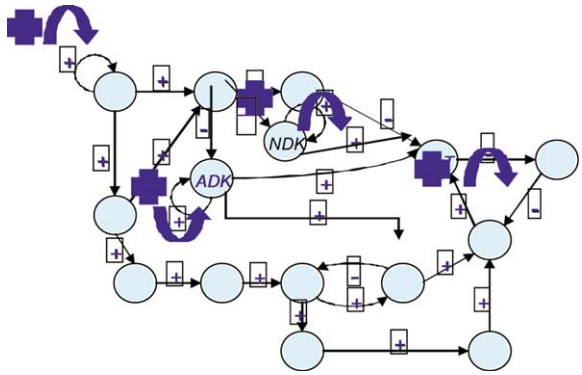


Fig. 6. The gastrulation regulatory network (after [27]) (ADK and NDK are respectively the adenylate kinase and the nucleotide diphosphate kinase).

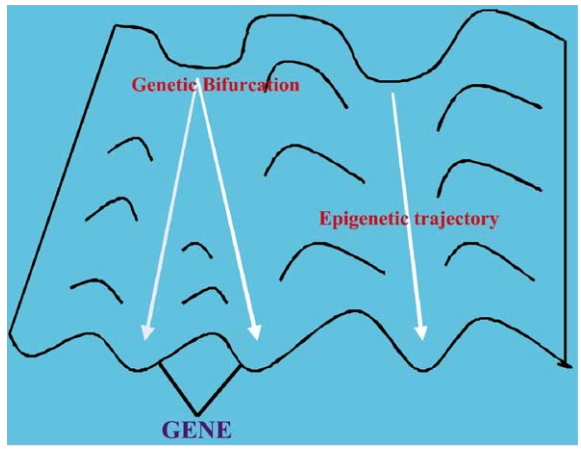


Fig. 7. The Waddington chreode or morphogenetic landscape.

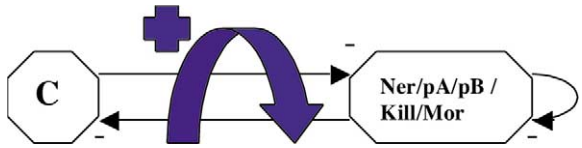


Fig. 8. The phage μ operon.

process due to the genetic regulation pathway of Fig. 6. This context describing the morphogenesis from genetic and epigenetic forces was called by Waddington a chreode or a morphogenetic landscape (cf. Fig. 7).

3.3. The phage μ lytic-lysogenic attractor [32]

If we consider the operon governing the expression of the phage μ, we obtain the graph given in Fig. 8. It is interesting to notice that $K(\mathbf{W}) = 3/2$, $I(\mathbf{W}) = 1$, $P(\mathbf{W}) = 1$, $A(\mathbf{W}) = 2$ (the two corresponding states in the host cell being the lytic and lysogenic ones, like for the Cro operon of the phage λ [24]). There is only one connected component.

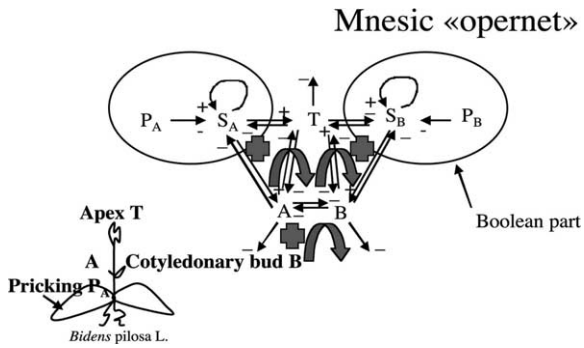


Fig. 9. The mnesic opernet.

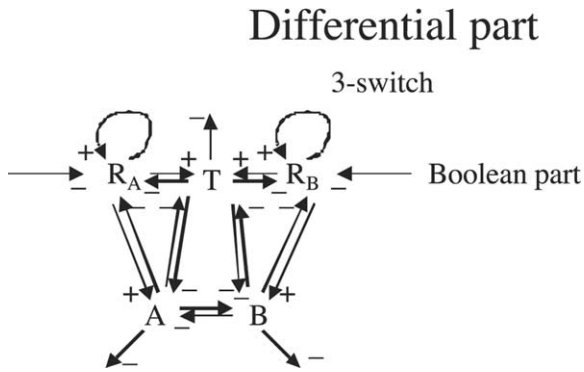


Fig. 10. Interaction graph of the epigenetic part as a continuous differential system.

3.4. The mnesic 'opernet'

We will call in the following mnesic 'opernet' the system obtained by merging the genetic (*oper*on) and epigenetic parts (metabolic *net*) of the system ruling the cotyledonary buds growth (cf. [33,34] and Fig. 9).

The genetic part of the mnesic opernet has been modelled by a Boolean system [34]:

- variable P_A (resp P_B) represents pricking treatment (or any other stress action) on the side A (resp B) of the plant and its value is 1 if treatment has been done, and 0 if not;
- variable S_A (resp S_B) represents the discrete part of the operon; we suppose that it contributes to mobilize a morphogenetic cotyledonary material R_A (resp R_B) responsible for the growth of the apex and of the cotyledonary bud A (resp B).

We will suppose in the following that the variable R_A (resp R_B) representing the concentration of R on side A (resp B) is continuous and that its velocity dR_A/dt (resp dR_B/dt) is ruled by a differential system containing a three-switch between the continuous variables T (apex growth metabolites concentration), A and B (cotyledonary buds A and B growth metabolites concentrations on respectively A and B side) (see Fig. 10).

Graph G of Fig. 9 is such that $K(\mathbf{W}) = 16/5$, $I(\mathbf{W}) = 10/16$, $P(\mathbf{W}) = 3$, $A(\mathbf{W}) = 3$; G is connected and contains two regulons.

Then we can write the differential system governing the continuous variables T, A, B, R_A and R_B as follows:

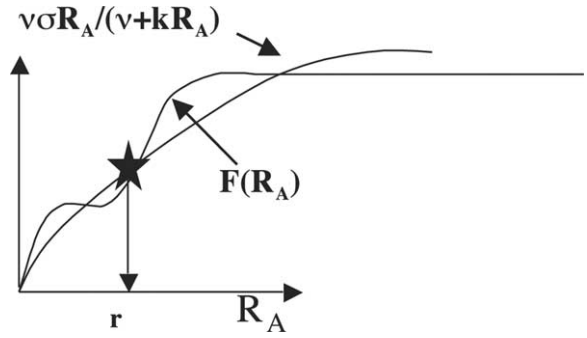


Fig. 11. Allosteric function F, which presents two successive inflection points.

$$
\begin{aligned}
\mathrm{d}R_\mathrm{A}/\mathrm{d}t &= (\sigma - kT - 4kA/5 - kB/5)R_\mathrm{A} - F(R_\mathrm{A})(T + 4A/5 + B/5)/(T + A + B) \\
&\quad - wP_\mathrm{A} - wP_\mathrm{B}/2 \\
\mathrm{d}R_\mathrm{B}/\mathrm{d}t &= (\sigma - kT - 4kB/5 - kA/5)R_\mathrm{B} - F(R_\mathrm{B})(T + A/5 + 4B/5)/(T + A + B) \\
&\quad - wP_\mathrm{B} - wP_\mathrm{A}/2 \\
\mathrm{d}T/\mathrm{d}t &= \big(F(R_\mathrm{A}) + F(R_\mathrm{B})\big)T/(T + A + B) - \nu T \\
\mathrm{d}A/\mathrm{d}t &= F(R_\mathrm{A})4A/5(T + A + B) + F(R_\mathrm{B})B/5(T + A + B) - \nu A \\
\mathrm{d}B/\mathrm{d}t &= F(R_\mathrm{B})4B/5(T + A + B) + F(R_\mathrm{A})A/5(T + A + B) - \nu B
\end{aligned}
$$

The two first equations correspond to the dynamics of R_A (resp R_B) whose concentration derivative at time $t\mathrm{d}R_\mathrm{A}(t)/\mathrm{d}t$ (resp $\mathrm{d}R_\mathrm{B}(t)/\mathrm{d}t$) results from an auto-catalytic term $\sigma R_\mathrm{A}(t)$ (resp $\sigma R_\mathrm{B}(t)$) diminished by the term $(-kT(t) - 4kA(t)/5 - kB(t)/5)\ R_\mathrm{A}(t)$ (resp $(-kT(t) - 4kB(t)/5 - kA(t)/5)\ R_\mathrm{B}(t)$) expressing the inhibition by T, A and B, plus a production of growth metabolites term denoted by $F(R_\mathrm{A}(t))\ (T(t) + 4A(t)/5 + B(t))/5)/(T(t) + A(t) + B(t))$ (resp $F(R_B(t))(T(t) + A(t)/5 + 4B(t)/5)/(T(t) + A(t) + B(t))$), by supposing that the R_A (resp R_B) consumption is competitively inhibited by the bud growth on its side A (resp B) and by the bud growth on the other side and by considering that K_ms and $K_{\min\ \mathrm{hib}}$s are equal to 1, plus the instantaneous perturbation $wP_\mathrm{A}(t)$ from its side A (resp B) and $wP_\mathrm{B}(t)/2$ from the other side B (resp A), the value of $P_\mathrm{A}(t)$ (resp $P_\mathrm{B}(t)$) being 1 if the pricking treatment occurs on A (resp B) at time $t = t_\mathrm{P}$, and 0 elsewhere.

The equations for the apex and cotyledonary buds growth metabolites concentrations just express that their production comes from R_A and R_B, with a competitive inhibition by the other sources of growth, plus a linear degradation term.

We suppose now for interpreting a minima the experimental results given in Tables 1 and 2 of [33] that F is an allosteric function of order 4 having two successive inflection points (involving that the protein catalysing the production of apex and buds growth metabolites has four catalytic subunits) as allowed by the Monod–Wyman–Changeux equation (see [35] and Fig. 11).

If the value of F' verifies $F(r) - rF'(0) < 0$ and $\nu < F(r)/r - F'(r) < 2\ (krF'(r))^{1/2} - \nu$, then the differential system above possesses at most 16 stationary states, whose 0 is a stable focus, two (respectively $(r, 0, \sigma r/(\nu + kr))$ and $(0, r, \sigma r/(\nu + kr))$ are unstable focuses surrounded by limit cycles α and β, and $\mathrm{C}(r, r, 2\sigma r/(\nu + kr))$ is an attractor (either stable focus, or limit cycle) as shown in Fig. 12, by supposing known the dynamics of the inhibitory three-switch [36] between T, A and B.

More generally, if we have an inhibitory n-switch between the A_i's verifying:

$$\mathrm{d}A_i/\mathrm{d}t = KA_i/\sum_{j=1,\ldots,n} A_j - \nu A_i$$

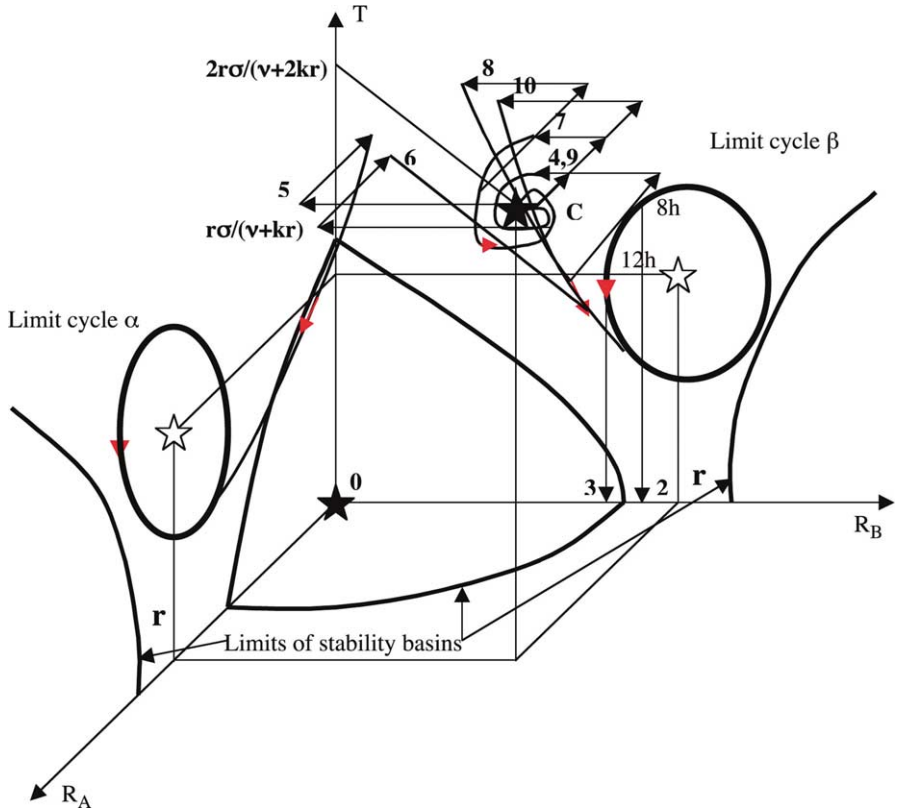


Fig. 12. Experimental perturbations in the state space (R_A, R_B, T).

then the stationary states $A_i = k = K/v$, $A_j = 0$, for $j \neq i$ are stable and the stationary states having $m > 1$ metabolites equal to $k = K/mv > 0$ and the other vanishing are unstable. It is easily to check this property on the Jacobian matrix of the differential system above [37] whose unique eigenvalue $\lambda = -v < 0$ in the first case and, in the second case, whose spectrum has the general eigenvalue $\lambda = -v + v(m-1)/m - rv/m - r^2v/m - \cdots - r^{m-1}v/m$, where r is one of the mth root of the unity. Then $\lambda = -(v/m)(1 + r + r^2 + \cdots + r^{m-1}) = 0$, if $r = e^{i2\pi/m}$, which implies the non-stability.

The possible configurations of experimental perturbations as reported in Table 1 below and in Tables 1 and 2 of [33] give different trajectories after perturbations, as explained in the following. If n_A (resp n_B) represents the number of seedlings beginning to elongate on the side of bud A (resp B), then $g = (n_A - n_B)/n$, where $n = n_A + n_B$, is an asymmetry growth index.

We can make in Fig. 12 above the following observations.

(1) If the initial conditions are inside the basin of stability of the attractor C, near C, then the whole trajectory without perturbation lies in this basin and tends to the attractor C (where $R_A \approx R_B$) as time t is tending to infinity. After decapitation without primitive pricking (non-pricked control), A and B buds have the same chance to begin to grow up, then to inhibit the bud growth on the opposite side, hence $g \approx 0$.

(2–3) If we do four pricking treatments on A, then from initial conditions near C we observe a shift to the basin of the limit cycle β and following the decapitation time, we get a B domination if decapitation is done at the onset daylight (dod) and a symmetry in buds growth if it occurs at midday (dm) (see Fig. 12): it is due to the

Table 1
Experimental results about domination growth after decapitation

Pricking treatment	g	Domination
(1) Non-pricked control	0.02	$A = B$
(2) 4A with decapitation at onset daylight (dod)	0.35 ± 0.02	$A < B$
(3) 4A with decapitation at midday (dm)	0.08 ± 0.15	$A = B$
(4) 1A (dod)	0.01	$A = B$
(5) 4B (dod)	-0.35	$A > B$
(6) 1A (1 h) 4B (dod)	0.39	$A < B$
(7) 2A (dod)	0.06	$A = B$
(8) 2A (1 h) 2A/2B (dod)	0.32	$A < B$
(9) 2A (1 h) 2A/2B (3 h) 2A/2B (dod)	0.05	$A = B$
(10) 2A (1 h) 2A/2B (3 h) 2A/2B (5 h) 2A/2B (dod)	0.34	$A < B$

fact that R_A and R_B after decapitation lies in the basin of the limit cycle β where $R_B > R_A$ in the first case and are in the basin of the stable focus 0 in the second case.

(4) If the system is starting near C and if one pricking 1A is done on cotyledon A (resp B), then we suppose that the perturbation results in a decrease of intensity w of R_A (resp R_B) and in a decrease of intensity $w/2$ of R_B (resp R_A) at time t_P of pricking; then the variables R_A and R_B remain in the basin of C.

(5) If we are doing four pricking treatments on side B, then the point representing the system leaves the basin of C with a decrease of $4w$ in R_A and $2w$ in R_B to go to the basin of the limit cycle β.

(6) If we do one pricking on side A and 1 h after four prickings on side B, then the system goes in the basin of the limit cycle α and for opposite reasons than for (5) above, $R_B < R_A$ and A dominates B after decapitation.

(7) If we do two prickings on side A, then the system remains in the basin of C.

(8) If we do two prickings on side A and 1 h after two prickings on each side, then the system goes in the basin of the limit cycle β.

(9) If we do two prickings on side A, 1 h after two prickings on each side and 3 h after two prickings on each side, then the system returns in the basin of C due to the special shape of the trajectory (8) going to the limit cycle β passing near the frontiers of the basin of C.

(10) If we do two prickings on side A, 1 h after two prickings on each side, 3 h after two prickings on each side and 5 h after anew two prickings on each side, the system goes in the basin of C like in (9) after 4 h, then 5 h after it goes from C to the basin of the limit cycle β.

We have then shown only by using the qualitative description of both the genetic (after pricking or any stress) and epigenetic forces exerted on the opernet that all the simulated behaviours described above qualitatively fitted the observed phenomenology (Table 1 above and Tables 1 and 2 of [33]).

4. Conclusion

An important first conclusion we have to make explicit in this paper concerns the relationship between the number F of fixed points and the number S of interaction circuits of the interaction matrix **W**: the problem is in fact to find the best upper bound for F for a given interaction matrix **W**. This question is related to the famous 16th Hilbert's problem, whose one of the aim is to give an efficient upper bound for the number of limit cycles of a polynomial differential system. Let us summarize the role of the architecture of positive and negative circuits of **W** on the occurrence of multiple stationary behaviours, as obtained above: if the number of nodes and the number of arcs are the same, there is only one isolated interaction circuit ($S = 1$) in **W** and either this circuit is negative and the lowest bound (0) for F is reached, or this circuit is positive and the upper bound (2^1) for F is reached. If the number of nodes is n and the number of arcs is $n + 1$, there is two interaction circuits ($S = 2$) with the

following structure: if both circuits are negative, $F = 0$; if there is a positive circuit and a negative circuit disjoint, $F = 0$; if there is a positive circuit intersecting a negative circuit, $F = 1$; if there is a positive circuit intersecting a positive circuit, $F = 1$; if there is two disjoint positive circuits, $F = 2^2$. If, more generally, the number S of interaction circuits of $\mathbf{W}$ is m, then: if all circuits are negative, $F = 0$; if all circuits are positive, $2 \leqslant F \leqslant 2^m$ and if all circuits are positive and disjoint, $F = 2^m$. An interesting open problem is now to make exhaustive the determination of F and S and in particular to find the circumstances (related to the circuits structure) in which we can relate the number of intersecting and isolated circuits to F. A conjecture we could make is that $F = 2^c$, where c is the number of not-singleton connected components of the interaction graph G having at least one positive loop: it holds for the lactose, *Arabidopsis*, phage μ and gastrulation regulatory systems. The approach for solving this open problem could consist in finding coherent relationships between analogous properties discovered independently for continuous and Boolean versions of regulatory networks [4–9].

The second conclusion concerns the practical use of the presented results; a geneticist can for example exploit the minimality results in the following sense: we have shown in the paper that it would be possible to characterize the minimal interaction matrices having certain state vectors as fixed points. The determination of these matrices is not unique, but permits to focus on a certain important equivalence class to which the expected matrix has to belong. This considerably restricts the choice of the possible interaction matrices compatible with observed fixed points, when it is impossible to directly get from experiments all interaction coefficients, but when it is only possible to observe the phenomenology of fixed points or limit cycles. This corresponds in genetics to the observation of stationary expression behaviours (for example from bio-array imaging) without experimental measure of the inhibitory and activatory coefficients of repressors and promoters. The possibility to obtain (even in an equivalence class) a sketch of the interaction matrix permits to construct (by randomising in a Bayesian way the unknown coefficients of $\mathbf{W}$) more complicated interaction matrices, then to test if they still have the observed states as fixed points and finally keep or reject definitively the so-tested matrices and propose further experimental strategies refining the knowledge about the interaction structure of a genetic regulatory network and then answer crucial biological questions like the relationship between genetic expression and recombination [16,38–40] (the crossing-over and translocations break points seeming correlated with the ubiquitory genes expression sites) or the relative parts taken by genetic and epigenetic forces in morphogenesis (embryogenesis or tumorogenesis). The last (but not the least) application of the interaction matrices introduced above is the ability to calculate the barycentre between two matrices by using classical (spectral or L_2) distances between matrices, then we could build phylogenic trees among a set of species avoiding the complex problems coming from the non-unicity of L_1 (Hamming or Manhattan) barycentres met in the sequence based phylogenic trees. The interaction based phylogenic trees could reflect more the genomic function than the genomic anatomy hence could explain more deeply the evolution trends.

Acknowledgements

We have done this work thanks to the support of the National Network for Technology Research RNTS ‘Technologies for Health’ from the French Ministry of Research.

References

[1] J. Demongeot, J.-P. Françoise, M. Richard, F. Senegas, T.P. Baum, A differential geometry approach for biomedical image processing, C. R. Biologies 325 (2002) 367–374.
[2] J. Demongeot, M. Richard, New segmenting and matching algorithms as tools for modeling and comparing medical images, Imacs 2000, EPFL, Lausanne, 2000, CD 127.
[3] J. Mattes, M. Richard, J. Demongeot, Tree representation for image matching and object recognition, Lect. Notes Comput. Sci. 1568 (1999) 298–309.
[4] E. Plahte, T. Mestl, S.W. Omholt, Feedback loops, stability and multi-stationarity in dynamical systems, J. Biol. Syst. 3 (1995) 409–414.

[5] J. Demongeot, Multi-stationarity and cell differentiation, J. Biol. Syst. 6 (1998) 1–2.
[6] E.H. Snoussi, Necessary condition for multi-stationarity and stable periodicity, J. Biol. Syst. 6 (1998) 3–10.
[7] J.-L. Gouzé, Positive and negative circuits in dynamical systems, J. Biol. Syst. 6 (1998) 11–16.
[8] O. Cinquin, J. Demongeot, Positive and negative feedback: striking a balance between necessary antagonists, J. Theoret. Biol. 216 (2002) 239–246.
[9] O. Cinquin, J. Demongeot, Positive and negative feedback: mending the ways of sloppy systems, C. R. Biologies 325 (2002) 1085–1095.
[10] L. Mendoza, E.R. Alvarez-Buylla, Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis, J. Theoret. Biol. 193 (1998) 307–319.
[11] E.H. Snoussi, R. Thomas, Logical identification of all steady states: the concept of feedback loop characteristic states, Bull. Math. Biol. 55 (1993) 973–991.
[12] J. Demongeot, M. Kaufman, R. Thomas, Interaction matrices, regulation circuits and memory, C. R. Acad. Sci. Paris, Ser. III 323 (2000) 69–80.
[13] J. Demongeot, J. Aracena, S. Ben Lamine, M.-A. Mermet, O. Cohen, Hot spots in chromosomal breakage: from description to etiology, in: D. Sankoff, J.H. Nadeau (Eds.), Comparative Genomics, Kluwer, Amsterdam, 2000, pp. 71–85.
[14] J. Demongeot, J. Aracena, S. Ben Lamine, S. Meignen, A. Tonnelier, R. Thomas, Dynamical systems and biological regulations, in: E. Goles, S. Martinez (Eds.), Complex Systems, Kluwer, Amsterdam, 2000, pp. 107–151.
[15] J. Aracena, J. Demongeot, E. Goles, Fixed points and maximal independent sets on AND-OR networks, Discrete Appl. Math. (in press).
[16] J. Aracena, S. Ben Lamine, M.-A. Mermet, O. Cohen, J. Demongeot, Mathematical modelling in genetic networks: relationships between the genetic expression and both chromosomic breakage and positive circuits, in: N. Bourbakis (Ed.), BIBE 2000, IEEE, Piscataway, 2000, pp. 141–149.
[17] C. Berge, Graphes et Hypergraphes, Dunod, Paris, 1974.
[18] E. Goles, S. Martinez, Neural and Automata Networks, in: Maths. Appl. Ser., Vol. 58, Kluwer, Amsterdam, 1991.
[19] B. Bollobas, Random Graphs, Academic Press, London, 1985.
[20] S. Kauffman, The Origins of Order, Oxford University Press, Oxford, UK, 1993.
[21] R. Thomas, On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations, Springer Ser. Synerget. 9 (1980) 1–23.
[22] M. Delbrück, Discussion, Unités biologiques douées de continuité génétique, Colloques internationaux CNRS 8 (1949) 33–35.
[23] R. Thomas, D. Thieffry, M. Kaufman, Dynamical behavior of biological regulatory networks. I. Biological role and logical analysis of feedback loops, Bull. Math. Biol. 57 (1995) 328–339.
[24] D. Thieffry, M. Colet, R. Thomas, Formalization of regulatory networks: a logical method and its automatization, Math. Model. Sci. Comput. 2 (1993) 144–151.
[25] R. Thomas, R. D'Ari, Biological Feedback, CRC Press, Boca Raton, 1990.
[26] F. Plouraboué, H. Atlan, G. Weisbuch, J.-P. Nadal, A network model of the coupling of ion channels with secondary messenger in cell signaling, Network Computation in Neural Networks Systems 3 (1992) 393–406.
[27] J. Aracena, Modèles mathématiques discrets associés à des systèmes biologiques. Application aux réseaux de régulation génétique, PhD thesis, U. Chile & UJF, Santiago, Chile, & Grenoble, France, 2001.
[28] M. Leptin, Gastrulation in *Drosophila*: the logic and the cellular mechanisms, EMBO J. 18 (1999) 3187–3192.
[29] N. Rashevsky, Mathematical Biophysics, Cambridge United Press, London, 1948.
[30] A. Turing, The mathematical basis of morphogenesis, Phil. Trans. Ro. Soc. B 237 (1952) 37–47.
[31] A.N. Kolmogorov, I. Petrowski, N. Piscounov, Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique, Mosc. Univ. Bull. Math. 1 (1937) 1–25.
[32] F. Thuderoz, DEA Report, UJF, Grenoble, France, 2000.
[33] M. Thellier, L. Le Sceller, V. Norris, M.C. Verdus, C. Ripoll, Long-distance transport, storage and recall of morphogenetic information in plants. The existence of a sort of primitive plant 'memory', C. R. Acad. Sci. Paris, Ser. III 323 (2000) 81–91.
[34] J. Demongeot, M. Thellier, R. Thomas, A mathematical model for storage and recall functions in plants, C. R. Acad. Sci. Paris, Ser. III 323 (2000) 93–97.
[35] J. Demongeot, M. Laurent, Sigmoidicity in allosteric models, Math. Biosci. 67 (1983) 1–17.
[36] R. Thomas, La logique des systèmes vivants, Bull. Cl. Sci. Acad. R. Belg. 74 (1988) 432–442.
[37] O. Cinquin, J. Demongeot, Inhibitory n-switch dynamics and applications, Math. Biosci. (in preparation).
[38] O. Cohen, M.A. Mermet, J. Demongeot, HC Forum®: a web site based on an international human cytogenetic data base, Nuclear Acids Research 29 (2001) 305–307.
[39] O. Cohen, C. Cans, M. Cuillel, J.-L. Gilardi, H. Roth, M.A. Mermet, P. Jalbert, J. Demongeot, Cartographic study: breakpoints in 1574 families carrying human reciprocal translocations, Hum. Genet. 97 (1996) 659–667.
[40] O. Cohen, C. Cans, M.-A. Mermet, J. Demongeot, P. Jalbert, Viability thresholds for partial trisomies and monosomies. A study of 1159 viable unbalanced reciprocal translocations, Hum. Genet. 93 (1994) 188–194.

Annex 3

Bio-array images processing and genetic networks modeling. Demongeot J, Thuderoz F, Baum TP, Berger F, Cohen O. C. R. Acad. Sci. Biologies. 2003. 326:487-500.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 326 (2003) 487–500
COMPTES RENDUS BIOLOGIES

Biological modelling / Biomodélisation

Bio-array images processing and genetic networks modelling

Jacques Demongeot [a,*,1], Florence Thuderoz [a], Thierry Pascal Baum [a], François Berger [b], Olivier Cohen [a]

[a] *TIMC–IMAG, CNRS 5525, Faculty of Medicine, 38700 La Tronche, France*
[b] *INSERM U 318, University Hospital of Grenoble, 38700 La Tronche, France*

Received 27 May 2002; accepted 4 March 2003

Presented by Michel Thellier

Abstract

The new tools available for gene expression studies are essentially the bio-array methods using a large variety of physical detectors (isotopes, fluorescent markers, ultrasounds...). Here we present first rapidly an image-processing method independent of the detector type, dealing with the noise and with the peaks overlapping, the peaks revealing the detector activity (isotopic in the presented example), correlated with the gene expression. After this primary step of bio-array image processing, we can extract information about causal influence (activation or inhibition) a gene can exert on other genes, leading to clusters of genes co-expression in which we extract an interaction matrix **M** and an associated interaction graph G explaining the genetic regulatory dynamics correlated to the studied tissue function. We give two examples of such interaction matrices and graphs (the flowering genetic regulatory network of Arabidopsis thaliana and the lytic/lysogenic operon of the phage Mu) and after some theoretical rigorous results recently obtained concerning the asymptotic states generated by the genetic networks having a given interaction matrix and reciprocally concerning the minimal (in the sense of having a minimal number of non-zero coefficients) matrices having given stationary stable states. ***To cite this article: J. Demongeot et al., C. R. Biologies 326 (2003).***

Résumé

Traitement d'images bio-array et modélisation de réseaux génétiques. Cet article décrit d'abord rapidement quels sont les nouveaux outils utilisés pour étudier l'expression des gènes, essentiellement les bio-arrays, qui mettent en œuvre un grand nombre de détecteurs physiques (isotopes, marqueurs fluorescents, ultra-sons...). Nous présentons une méthode de traitement d'images indépendante du type de détecteur, traitant le problème du bruit et des superpositions de pics, ces derniers révélant l'activité du détecteur (isotopique dans le cas choisi ici) corrélée avec l'expression des gènes correspondants. Après ce premier stade de traitement d'images bio-array, on peut extraire l'information relative à l'influence (activation ou inhibition) qu'un gène peut exercer sur les autres gènes, conduisant ainsi à l'apparition de groupes de co-expression, d'où l'on peut extraire une matrice d'interaction **M** et un graphe d'interaction associé G, susceptibles d'expliquer la dynamique de la régulation génétique corrélée avec la fonction tissulaire associée. Nous donnons quelques exemples de telles matrices et de tels graphes d'interaction (en particulier dans le cas du réseau de régulation génétique de la floraison d'*Arabidopsis thaliana* et dans celui de l'opéron lytique/lysogénique du phage Mu), et ensuite quelques résultats théoriques rigoureux récemment obtenus sur les états

* Corresponding author.
E-mail address: Jacques.Demongeot@imag.fr (J. Demongeot).
[1] Institut universitaire de France.

1631-0691/03/$ – see front matter
doi:10.1016/S1631-0691(03)00114-8

asymptotiques générés par des réseaux génétiques ayant une matrice d'interaction donnée. Réciproquement, nous décrirons les matrices minimales (au sens du nombre de leurs coefficients non nuls) ayant des états stationnaires stables donnés. ***Pour citer cet article : J. Demongeot et al., C. R. Biologies 326 (2003).***

Keywords: bio-array images; genetic network; inter-genic interaction matrix; positive loops; operons

Mots-clés : images *bio-array* ; réseau génétique ; matrice d'interaction inter-génique ; boucles positives ; opérons

1. Introduction

The total mRNAs of genes to test are extracted from the studied tissue (in the present case a glioblastome tissue). DNAs are synthesized by reverse transcription from these mRNAs including bases labelled with the isotope P^{33}. Resulting DNAs are then tested against identified complementary DNAs (cDNA targets), previously amplified by PCR and fixed on a nylon gel. The hybridisation results are revealed in a phosphoimager and yield a digital image coming from the radioactive hybridisation plate, called the bio-array image or shortly the bio-image. cDNA hybridised with a P^{33}DNA means that the complementary sequence of the P^{33}DNA is present in the related mRNA, proving that the corresponding gene is expressed in the studied tissue.

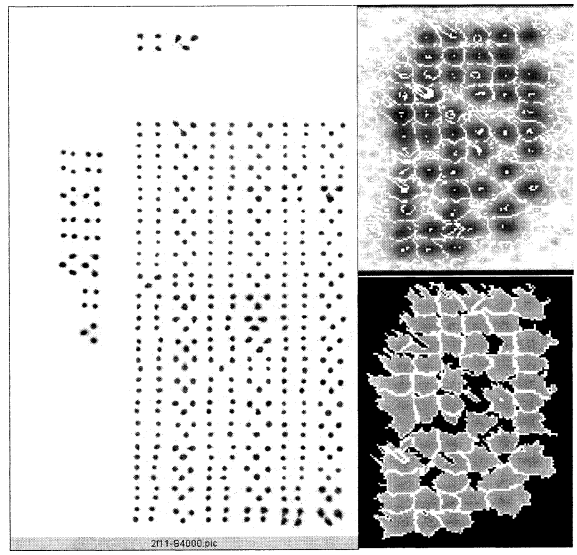


Fig. 1. Raw data (left), watershed segmenting (top right), and contouring (bottom right).

2. Peak segmentation

The first encountered problem is the fact that the bio-images are extremely noisy and that we have to low-pass them in order to suppress the high-frequency noise. The result of this pre-treatment (Fig. 1) is a better separation of the isotopic activity peaks, allowing a watershed separation and contouring [1,2], but it often leads to over-estimating the peak activity.

Then we will apply a more accurate segmentation and contouring method called the potential-Hamiltonian or 'Gaussian stamping' method: let us remark that the peaks are about Gaussian, with a relatively weak kurtosis and skewness allowing in particular the respect of the conservation 'law': 2/3 of the peak activity are concentrated into the set of points (x, y) where the Gaussian curvature $H(x, y)$ vanishes, i.e. inside the maximum gradient line of the peak. By exploiting this property, it is possible to neglect the part of the peak outside the projection of this remarkable line, called in the following the characteristic line, its equation being:

$$H(x, y) = \partial^2 g/\partial x^2\, \partial^2 g/\partial y^2 - (\partial^2 g/\partial x\, \partial y)^2 = 0$$

where $g(x, y)$ is the grey function at the pixel (x, y).

We are thus led to consider the new grey function $H(x, y)$ instead of the function $g(x, y)$ and its level line $H(x, y) = 0$. We display after a plane differential system of which the characteristic line is a limit cycle. Let $H'(x, y)$ be the function defined by: $H'(x, y) = |H(x, y)|$. Vanishing of $H'(x, y)$ occurs on the characteristic line (see Fig. 2 for the visualization of g and H') and if we consider the following crude system:

$$\mathrm{d}x/\mathrm{d}t = -\alpha\, \partial H'/\partial x + \beta\, \partial H'/\partial y$$
$$\mathrm{d}y/\mathrm{d}t = -\alpha\, \partial H'/\partial y - \beta\, \partial H'/\partial x$$

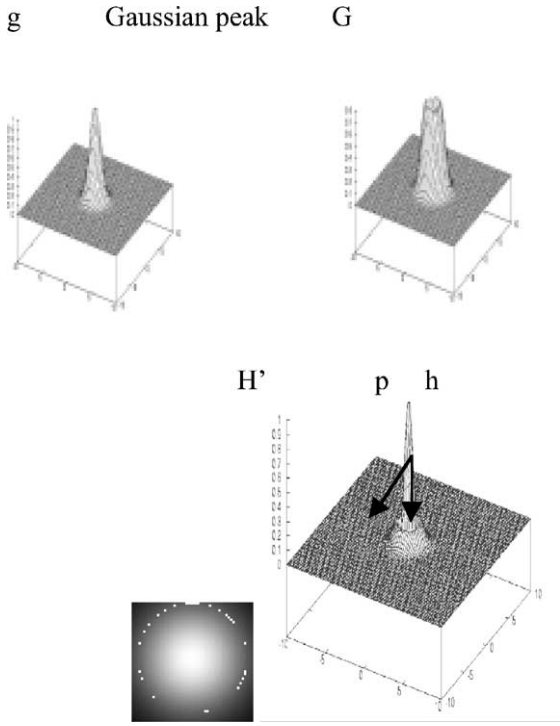


Fig. 2. Representation of g, G, H' (with indication of potential p and Hamiltonian h parts) for a Gaussian peak and the result of the Hamiltonian segmentation (below left).

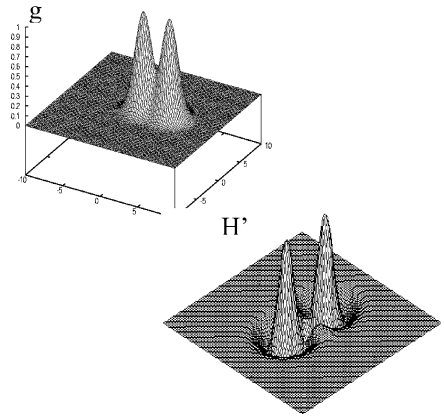


Fig. 3. g (top) and H' (bottom) for close Gaussian peaks.

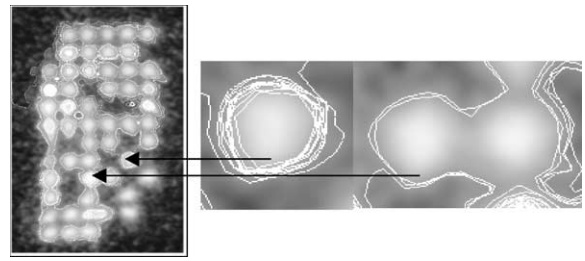


Fig. 4. Treated bio-image (left), succeeding limit cycle (middle) and false contour (right).

where α and β are real parameters, then the first part of this differential system is of steepest descent potential nature and along this flow, the orbits converge to the set of zeros of $H'(x, y)$, on which the second part of convective Hamiltonian type becomes preponderant [1]. Parameters α and β are used to tune the speed of convergence to the limit cycle. To cope with random noise and numeric instabilities, we modify slightly the system into:

$$\mathrm{d}x/\mathrm{d}t = -\alpha\,\partial H'/\partial x\big[H(x,y)/G(x,y)\big] + \beta\,\partial H'/\partial y$$
$$\mathrm{d}y/\mathrm{d}t = -\alpha\,\partial H'/\partial y\big[H(x,y)/G(x,y)\big] - \beta\,\partial H'/\partial x,$$
$$\text{where } G(x,y) = \big\|\mathrm{grad}(g)\big\|^2$$

The added term $H(x, y)/G(x, y)$ speeds up the descent to the vanishing of $H(x, y)$ and forces the stability. The usual discretization of Runge–Kutta yields ultimately the algorithm, which is quite easy to implement. On each pixel (i, j) – boundary effects being neglected –, the function $H(i, j)$ reads:

$$H(i,j) = \big[g(i+2,j) - 2\,g(i+1,j) + g(i,j)\big]$$
$$\times\big[g(i,j+2) - 2\,g(i,j+1) + g(i,j)\big]$$
$$-\big[g(i+1,j+1) - g(i,j+1) - g(i+1,j) + g(i,j)\big]^2$$

We have seen an important property of the characteristic line, i.e. in the case of a Gaussian peak, it delimits a volume equal to $2/3$ of the total volume of the peak. This property remains about exact in case of kurtosis and skewness of the peak. Hence by multiplying by $3/2$ this volume, we get a good estimation of the gene activity and this value is better than that obtained by a watershed method due to over-segmentation (Fig. 1). This approach is interesting because the lower part of the peak is often noisy. The method seems particularly efficient when the peaks are well separated. If they are close (Fig. 3), then we need to tune the parameters α and β (Fig. 4). In further developments of the method, we look for a dynamical calculation of these parameters from the data. Finally, we can standardize the estimated activity in terms of a bio-image with small squares symbolizing in grey levels the de-

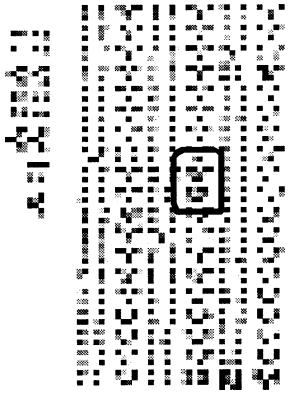


Fig. 5. Standardized bio-image (the treated sub-image of Fig. 4 is inside the black rectangle).

gree of hybridisation of the cDNAs (Fig. 5). From such bio-images acquired at different times of the cell cycle in cells from the same tissue, we can study the interactions between genes by estimating an interaction matrix.

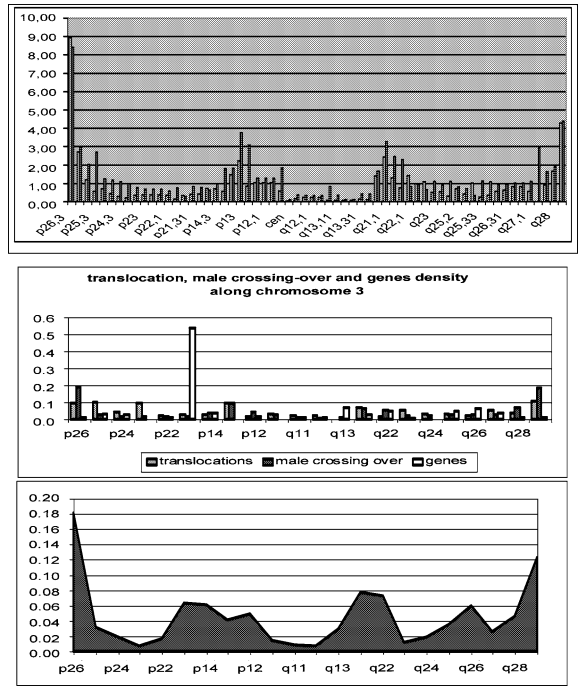


Fig. 6. Distributions of domestic (bottom) and all genes (middle right), crossing-over (male red and female blue top) and translocations (middle left) along chromosome 3.

3. Interaction matrix M

A major problem a geneticist has presently to face since the introduction of the bio-array imaging is the estimation of the intergenic interaction matrix **M** that rules the observed genes expression in operons and genetic regulatory networks [3–7]. This interaction matrix is similar to the synaptic weight matrix, which rules the relationships between neurons in a neural network. Hence, it is in general of a great biological interest and relevance to determine matrices having characteristics like: (*i*) a minimal number of non-zero coefficients for a given set of stationary behaviours (fixed points or cycles), (*ii*) a minimal number of positive or negative circuits, controlling the number of attractors and their stability (cf. [8–21] for both the discrete and the continuous case). In this paper, we give some general results about the relationships between the positive and negative circuits in the graph of the interaction matrix **M** and the existence of fixed points. This permits us to characterize minimal matrices, given dynamical behaviours, and therefore partly solve the first problem. Finally, we constructed a bound for the number of fixed points in terms of the number of positive circuits in the graph of the interaction matrix **M**. So we partly solve the second problem too.

In general, it is very difficult to have exhaustively the interaction matrices: in the genetic literature and also by observing co-expressions through bio-arrays imaging, it is possible to qualitatively or even quantitatively estimate the inhibitory (in case of repression by a protein obtained from the expression of a gene) or activatory (in case of induction or promotion) coefficients of the interaction matrix. If we have no information, we can randomly choose the matrix by respecting certain basic rules, e.g., by respecting certain proportions of activatory or inhibitory interactions. We can, for example, obtain the location density of expressed ubiquitory genes (calculated from http://www.citi2.fr/GENATLAS/) and then randomly simulate the interaction matrix and the initial conditions of the gene expression, by sampling them 100 000 times, the interaction matrices respecting the constraint to have 10% (resp. 10%) of negative (resp. positive) interactions, like in the *Arabidop-*

Table 1
Second column shows the distribution signatures and third the number of differences (− or + being not taken different from =) between a given distribution and the crossing-over one as reference

Ubiquitory genes distribution	−	−	−	+	+	=	−	+	−	=	=	+	+	=	−	+	+	+	−	+	+	3/21
All genes distribution	+	=	−	+	+	−	−	+	−	+	=	+	−	+	−	=	+	+	−	−	+	9/21
Crossing-over distribution	−	=	=	=	=	+	+	−	−	−	=	+	+	−	−	=	=	=	+	+	+	0/21
Translocation distribution	=	−	+	−	=	=	+	−	+	−	=	=	+	−	+	−	=	=	+	−	+	3/21

sis thaliana genome [3]. Fig. 6 below gives the distribution of the co-expression of the ubiquitory genes calculated from the expected stationary behaviour corresponding to a random choice of the interaction matrix and of the initial conditions: we have systematically calculated attractors (fixed points or limit cycles) corresponding to an initial condition and an interaction matrix, and then we have calculated the frequency of observing the expression of each ubiquitory gene in these attractors. In the absence of complementary information about the localization of the inhibitory or activatory interactions between ubiquitory genes, the obtained co-expression distribution is just a reflect of the spatial distribution of these ubiquitory (domestic or housekeeping) genes along the human chromosome 3 showing that it is related to the rearrangements (due to physiological crossing-over or to pathological translocations) distributions, this parenthood being proved by the analogy between their signatures—i.e. their succession of monotony intervals of increase (+) or decrease (−)—as shown in Table 1.

By comparing the distribution signatures given in Table 1 above, it is easy to prove that we must reject the hypothesis that ubiquitory genes and translocation distributions are different from the crossing-over distribution ($p < 0.001$), but we cannot reject the hypothesis of a difference between all genes and the crossing-over distribution.

The general coefficient m_{ij} of the interaction matrix $\mathbf{M}$ is equal to $+1$ if the gene G_j activates the gene G_i, equal to -1 if the gene G_j inhibits the gene G_i and equal to 0 if G_j and G_i have no interaction, G_i being equal to $+1$ (resp. -1), if it is (resp. not) expressed. Then the change of state x_i of the gene G_i between t and $t+1$ obeys a threshold rule: $x_i(t+1) = H(\sum_{k=1,n} m_{ik}x_k(t) - b_i)$ or $x(t+1) = H(\mathbf{M}x(t) - b)$, where H is the sign-step function ($H(y) = 1$, if $y \geqslant 0$ and $H(y) = -1$, if $y < 0$) and the b_is are threshold values. In the case of small regulatory genetic systems (the smallest being called operons), the knowledge of such a matrix $\mathbf{M}$ permits to make explicit all possible stationary behaviours of the organisms having the corresponding genome: for example, in the genetic regulatory network that rules the *Arabidopsis thaliana* flower morphogenesis (Fig. 7), the interaction matrix is a $(11, 11)$-matrix, with only 22 non-zero coefficients. This matrix presents a certain number of positive or negative circuits and only four observed attractors [3].

For each operon, we can define an interaction matrix $\mathbf{M}$, which just expresses that if its coefficient m_{ij} is positive (resp. negative), the gene j is a promoter or activator (resp. repressor or inhibitor) of the gene i. If m_{ij} is null, then the gene G_j has no influence on the expression of the gene G_i. The interaction graph can be built from the interaction matrix $\mathbf{M}$ by drawing an edge + (resp. −) between the vertices representing the genes j and i, if $m_{ij} > 0$ (resp. < 0). In order to calculate the m_{ij}s, we can either determine the s-directional correlation $\rho_{ij}(s)$ between the state vector $\{x_j(t-s)\}_{t\in C,\, t\geqslant s}$ of gene j at time $t-s$ and the state vector $\{x_j(t)\}_{t\in C,\, t\geqslant s}$ of gene i at times t, t varying during the cell cycle C, or identify the system with a Boolean neural network.

We define the connectivity $K(\mathbf{M})$ of the interaction matrix $\mathbf{M}$ by the ratio between the numbers m of edges of the interaction graph and n of vertices: in general, for known operons and genetic regulatory networks (lactose operon [5,6], Cro operon of the phage λ, lysogenic/lytic operon [4,7] of the phage infecting *E. coli*, gastrulation regulatory network...), $K(\mathbf{M})$ is between 1.5 and 3. The observed induction proportion (number of positive edges divided by m) is between 1/3 and 1/2. If m_{ij}s are unknown, we can take them randomly by respecting connectivity and induction proportion.

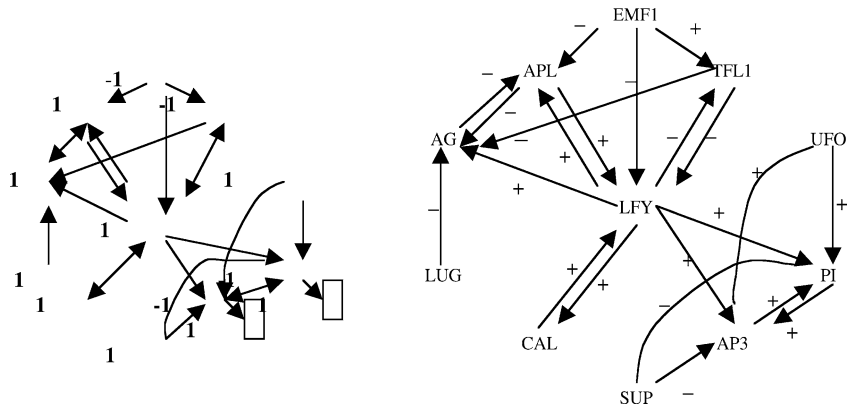


Fig. 7. Interaction graph of the flowering genetic regulatory network of *Arabidopsis thaliana* (right) and an attractor of its Boolean dynamics (left).

4. Genetic networks dynamics

If we consider the interaction graph of the flowering genetic regulatory network of *Arabidopsis thaliana* (Fig. 7) [3], then we can easily define from it a Boolean dynamics with threshold 0: the gene i has the state 1 if it is expressed and -1 if not. The change of state of gene i between the times t and $t+1$ obeys a majority rule, i.e. we calculate the numbers of its neighbours in state 1 with positive interaction and with negative interaction: if these two numbers are equal, then the new state of i is 1; if the activatory (resp. inhibitory) neighbours dominate, then the new state of i is 1 (resp. -1). When the time t is increasing, the configuration of gene states reaches a stable set of configuration (either a fixed configuration or a cycle of configurations), called an attractor of the genetic network dynamics. In Fig. 7 (left), an example of such an attractor is given, with final states (in black boxes) different from the initial conditions.

We will present now first some qualitative results from the human genome observation, and after some theoretical corresponding statements recently proved:

– in 1949, Delbrück [17] conjectured that the presence of positive loops (i.e. paths from a gene i to itself having an even number of inhibitions [11]) in the interaction graph was a necessary condition for the cell differentiation; this conjecture has been more precisely written in a good mathematical context by Thomas in 1980 [16];

– in 1993, Kauffman [22] conjectured that the mean number of attractors for a Boolean genetic network with n genes and with connectivity 2, was equal to $\sqrt{n}$. This conjecture is supported by real observations: we have about 30 000 genes in the human genome and about 200 different tissues, which can be considered as different attractors of the same dynamics. For *Arabidopsis thaliana*, $K(\mathbf{M}) = 22/11 = 2$ and there is $4 \approx \sqrt{11}$ different tissues (sepals, petals, stamens, carpels) [3] and for Cro operon [18] of phage λ, $K(\mathbf{M}) = 14/5 = 2.8$ and there is $2 \approx \sqrt{5}$ observed (lytic and lysogenic) attractors.

Recently [8–15], these conjectures have been partially proved.

Proposition 1. *If all loops of the interaction graph are positive, then there exists a state vector* $x = (x_1, \ldots, x_n)$ *in* $\{-1, 1\}^n$, *such that* x *and* $-x = (-x_1, \ldots, -x_n)$ *are fixed configurations of the network dynamics.*

Proposition 2. *If all loops of the interaction graph are negative, then there is no fixed configuration.*

Proposition 3. *Let a network having* n *genes and* n *interactions, then a necessary and sufficient condition of existence of a fixed configuration* x *is the existence of a positive loop and* $-x$ *is also fixed.*

Proposition 4. *Given a state vector* **x**, *the set of minimal matrices* **M** *having* **x** *as fixed configuration is given by the following conditions*:

(1) $m_{ij} = a_{ij}\, x_i\, x_j$, *where* $a_{ij} \geqslant 0$ *and, for all* i, *there exists* $j(i)$ *such that* $a_{ij(i)} > 0$;
(2) $-a_{ij(i)} < b_i \leqslant a_{ij(i)}$, *where* a_{ij}*s and* b_i*s are weights and thresholds of the genetic network* [14].

Proposition 5. *If* m *is the total number of positive loops* C, *then the number of fixed configurations is less than or equal to* 2^m, *and this upper bound is reached, if and only if for any positive loop* C, *there is no edge* $(x, y) \mid x \notin \mathrm{C},\ y \in \mathrm{C}$.

Proposition 6. *If the genetic network has* n *genes and* $2n$ *interactions, then the expectation of the number of its fixed configurations is* $\sqrt{n}$, *if* n *is sufficiently large* [15].

Proposition 7. *If the interaction graph* G *is a connected digraph without loops having* n *genes and let suppose that, for any* $i, b_i > 0$ *and all* m_{ij}*s are positive and verify either* $(\sum_j m_{ij} \geqslant b_i$ *and, for any* k, $\sum_{j \neq k} m_{ij} < b_i) = \mathrm{AND}$ *rule, or (for each* j, $m_{ij} \geqslant b_i) = \mathrm{OR}$ *rule, then the number of fixed configurations is* $2^{(n-1)/2}$ *for* n *odd, and* $2^{(n-2)/2} + 1$ *for* n *even* [14].

Numerous applications of the results above are possible in various regulatory systems [23–40], but we will focus here in the following on a very simple operon governing the choice between the lytic and lysogenic stationary states for *Escherichia coli* infected by the bacteriophage Mu.

5. An example of operon: the lytic/lysogenic operon of the bacteriophage Mu

Understanding the behaviour of the tempered bacteriophage Mu, considered as a transposon, constitutes a real progress in the knowledge of transposition mechanisms [41].

A Boolean model of the interactions between the expression products of the Mu genome shows the necessity for the removal of the auto-inhibition of the protein Ner (negative loop), which allows the constitutive expression of the phage promoter Pe (positive loop) to demonstrate a bi-stability [11,42]. The presence of the prophage state instability is necessary for the induction of prophages. A review of *Escherichia coli* factors influencing the Mu behaviour allowed us to propose the Integration Host Factor (IHF) and the Inversion Stimulating Factor (ISF) as being responsible for the removal of auto-inhibition and for the prophagic state instability. The modelling of their effects by a differential system clearly shows the same lytic/lysogenic proportions than those experimentally obtained during the exponential bacterial growth phase and the stationary bacterial growth phase.

These encouraging results demonstrate the necessity of quantitative measures for lytic/lysogenic proportions and intra-cellular concentrations of different *Escherichia coli* factors during the entire length of the growth cycle, in order to better understand the induction context of the bacteriophage Mu.

The bacteriophage Mu has been discovered in 1963 by L. Taylor [43]. During the infection step, Mu incorporates its DNA and proteins at random locations in the host bacterial chromosome [44]. That induces mutations and new auxotrophies and Mu belongs to the large family of transposable elements [45]. There is two developmental cycles for Mu: after integration in the bacterial chromosome, the Mu DNA is multiplied by a series of replicative transpositions, giving between 50 and 100 new viruses after lysis of the bacterial host (lytic cycle). A weak proportion of infected bacteria becomes lysogenic, the Mu DNA remaining inactive (lysogenic cycle). The induction rate represents the proportion of such lysogenic bacteria entering in the lytic cycle.

Mu infects enterobacteria and in particular *E. coli* [46] and the integration of its genome into the chromosome of these bacteria involves the phagic transposase pA and its activator pB [47] (cf. Fig. 8). In general, this integrated DNA is amplified through a series of replicative transpositions [48] (involving the transposase pA and its activator pB), then the late phagic functions are synthesized, leading to assembling numerous viral particles dispatched in the extracellular medium after the lysis of the host cell. Among a population of infected bacteria, a small proportion will give birth to lysogenic cells, in which the phagic DNA remains passive; this DNA can be replicated during the further mitoses of the lysogenic cell or can enter in a lytic cycle. After integration, the choice be-

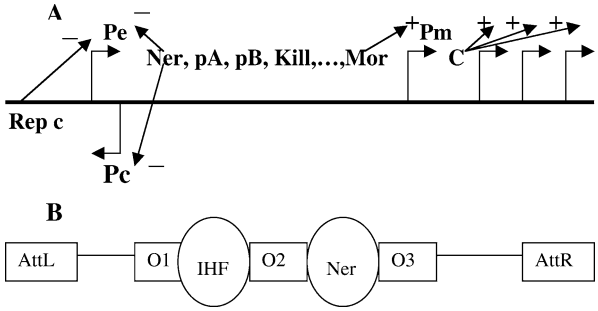


Fig. 8. Simplified scheme for the viral DNA of Mu; (A) shows different genes in transcriptional units and the actions of their expression products on the promoters' activity. The fixation sites of the different phagic (e.g., Ner on O_2) or bacterial (e.g., IHF on O_1) proteic factors are represented in (B).

tween the two possible developmental cycles (lytic or lysogenic) is regulated at the levels of transcription and replicative transposition of the Mu genome.

- *Transcriptional regulation.* The viral DNA has a size of 37 kb [49] and possesses two promoters Pe and Pc constitutively expressed [50] (Fig. 8A). The early transcript from Pe codes for the transposase pA, for its activator pB and for a protein Mor activating (via the promoter P and the protein C [51,52]) the cascade of the late functions of the lytic cycle. The transcript coming from Pc codes for the repressor c preferentially fixed on the operators O1 and O2 (Fig. 8B) for repressing Pe and stabilizing the Mu DNA in its inactive form (the prophage DNA). At high concentration, c is also fixed on O3, repressing Pc and hence regulating its own synthesis [53]. The Ner protein is fixed between Pe and Pc and inhibits the transcription from these two promoters [47,54].
- *Transpositional regulation.* During the transposition, the transposase pA transiently fix the extremities of the Mu DNA, attL and attR, and the operators O1 and O2 in stabilizing a tetrameric structure, which leads the Mu DNA to a specific configuration, the transposome. The repressor c binds also the extremities of the Mu DNA with a weak affinity [55], entering in competition with pA and then impeaching it to bind the two operators and the two extremities [56]. Hence repressor c inhibits the lytic cycle both at the transcriptional and transpositional levels, and the transposase pA being physically and functionally unstable in vivo [57], the viral genome amplification implies a non-transient expression of the promoter Pe.

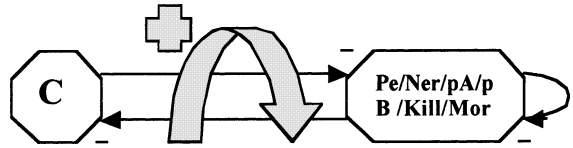
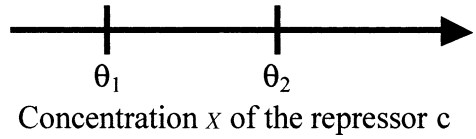


Fig. 9. Simplified scheme of the bacteriophage Mu operon (the blue arrow symbolizes the presence of a positive loop).

5.1. Discrete logical (Boolean) model

Experiments about the phage Mu are done during the exponential growth phase of the bacteria for increasing the homogeneity of the observed bacterial states. The lysogenic frequency is highly depending on the used protocol (1% to 50%). We will base our model on lysogenic induction rates observed in exponential phase (0.001%) and in stationary phase (about 100%) (cts 62, temperature 30 °C). For interpreting these experimental data, we used first Thomas' logical approach [16,18,20,21], in which x represents the repressor c and the entity y denotes the group of proteins Pe, Ner, pA, pB, Kill and Mor; their interactions can be designed as in Fig. 9.

Variables x and y take values 1 (expression/presence in the cell) or 0 (non-expression/absence from the cell). With this notation, the state E_x, $x = 1$ and $y = 0$ corresponds to the lysogenic state and the state E_y, $x = 0$ and $y = 1$ to the lytic cycle. Let us notice that it is not necessary to explicitly represent the auto-inhibition of the repressor c, because c inhibits Pe and auto-inhibits itself at high concentrations, hence maintaining its concentration between two thresholds θ_1 and θ_2:



Below θ_1 (e.g., $x = 0$), the repressor c has insufficient concentration to be fixed to the operators O1 and O2, hence x does not inhibit y. Between θ_1 and θ_2, x inhibits y but is not sufficient for fixing O3; hence there is no auto-inhibition of x. Above θ_2, x inhibits

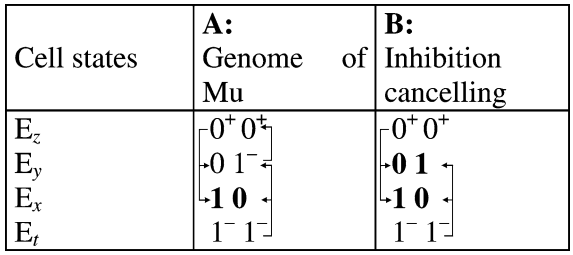


Fig. 10. Dynamical behaviour of Mu alone (A) and with a factor cancelling auto-inhibition of y (B). The stationary steady states are represented in bold and superscripts are corresponding to state changes.

y and inhibits itself until going to a level less than θ_2. Because this negative loop causes a homeostasis, we consider for x only the values $x = 0$ and $x = 1$, and the self-inhibition is not explicit in the model.

Because the two promoters Pe and Pc are constitutively expressed, x takes the value 1 if it is not inhibited by y: if $y = 0$, then $x = 1$ and if $y = 1$, then $x = 0$. Reciprocally, if $x = 0$, $y = 1$ and if $x = 1$, $y = 0$, leading to the following logic equations:

$$X = \neg y$$

$$Y = \neg x \times \neg y$$

where x and y represent the values of the two logic variables X and Y.

We consider the four possible initial states E_x ($x = 1$, $y = 0$), E_y ($x = 0$, $y = 1$), E_z ($x = 0$, $y = 0$) and E_t ($x = 1$, $y = 1$); if two variables 'enlightened' tend to become 'off' (noted 1^-) or conversely if two variables 'off' tend to become 'on' (noted 0^+), they are not changing at the same time. For example, 0^+0^+ gives 10 or 01. The behaviour of this system (cf. Fig. 10A) has only a stable stationary state (sss) E_x ($x = 1$, $y = 0$), which corresponds to the lysogenic state. But Mu is a bi-stable system and we have to put another information in the model to get a second sss.

For rendering stable the lytic state E_y ($x = 0$, $y = 1$), we need an activation of y cancelling its self-inhibition proportionally to its concentration [11]. In stationary phase, the proportion of lysogenes is near 100%, despite the fact that c is still present. Then we need two factors participating either to the maintenance of the passive phage state, or to the lytic cycle induction (both at the transcriptional and transpositional levels). A review of such possible cellular factors leads to the following list.

- IHF (Integration Host Factor). IHF is a histone which presents a high affinity for the operator part of the Mu DNA (Fig. 8B), increasing its curvature, hence facilitating the fixation of the repressor c on the operators [58,59] and activating the transcription from the promoter Pe passing over the retro-inhibition by Ner, if c is absent [60]. The absence of IHF impeaches any significative production of phage Mu [61]. The cell concentration of IHF is inversely proportional to the mitosis rate (from 12 000 mol/cell in exponential phase to 52 000 mol/cell in stationary phase [61]).
- HU (Histone U). It causes the same effects than IHF on transcription and transposition [62], but its concentration does not vary following the bacterial growth rate [61].
- ISF (Inversion Stimulating Factor). ISF helps c in repressing the lytic cycle: it increases the transcriptional repression and inhibits the transposition [63]; its absence multiplies by about 600 until 800 times the viral particles production in vivo [64]. Cell concentration of ISF is maximal at the start of the exponential growth phase (80 000 mol/cell) and decreases until a rate zero when the growth diminishes [65,66].
- 8-proteins system. For its DNA replication Mu depends also on host enzymes, from which eight have been identified [67] and for the degradation of the transposase pA and catabolism of c on 2 proteases [41].

5.2. Continuous differential model

IHF is the only bacterial factor activating the transcription of Pe in absence of c. For 01 (E_y) being a stable stationary state (sss), the activation of Pe by IHF has to equilibrate the repression by Ner, leaving the constitutive expression of Pe to run. In stationary growth phase, the induction of prophages is total, despite the presence of a measurable activity of c. The factor ISF helping the action of the repressor c, present in exponential phase but absent in stationary phase, is responsible for that behaviour. Let us now consider a continuous differential model, where the variables are the four states previously considered (E_x, E_y, E_t and E_z), IHF and ISF being parameters whose value changes can favour the passage from a state where one of these variables dominates to another state where it

is another dominating. The differential system can be written as:

$$\mathrm{d}E_x/\mathrm{d}t = -k_3\,E_x/\mathrm{ISF} + (1/3 + \mathrm{ISF}/k_2)\,E_t + (1/3 - \mathrm{IHF}/k_1)\,E_z$$

$$\mathrm{d}E_y/\mathrm{d}t = -k_4\,E_y/\mathrm{IHF} + (1/3 - \mathrm{ISF}/k_2)\,E_t + (1/3 + \mathrm{IHF}/k_1)\,E_z$$

$$\mathrm{d}E_t/\mathrm{d}t = -E_t + E_z/3$$

$$\mathrm{d}E_z/\mathrm{d}t = k_3\,E_x/\mathrm{ISF} + k_4\,E_y/\mathrm{IHF} + E_t/3 - E_z$$

We assume from experimental data that $k_3/k_4 = k_2/k_1 \sim 6$, where k_4 and k_1 are fixed in such a way that $E_y(\infty) = 10^5$ and $E_z(\infty) = 1$ (stationary phase condition). Then the non-zero steady state with $E_z(\infty) = 1$ verifies: $E_t(\infty) = 1/3$, $E_x(\infty) = \mathrm{ISF}(4/9 - \mathrm{IHF}/k_1 + \mathrm{ISF}/3k_2)/k_3$, $E_y(\infty) = \mathrm{IHF}(4/9 + \mathrm{IHF}/k_1 - \mathrm{ISF}/3k_2)/k_4$.

5.3. Results

We have simulated (Mapple Runge–Kutta 4 with constant steps) the dynamical behaviour of 100 000 phages Mu initially in lysogenic state ($E_x(0) = 10^5$, and $E_y(0) = E_z(0) = E_t(0) = 0$), in exponential growth phase (IHF = 12 000 mol/cell, ISF = 80 000 mol/cell) and in stationary growth phase (IHF = 52 000 mol/cell, ISF = 1 mol/cell) with parameters values equal to $k_1 = 100$, $k_2 = 700$, $k_3 = 1800$, $k_4 = 300$.

In stationary phase (Fig. 11), the lytic state (E_y) is increasing and reaching a plateau of 100 000 lytic cells, which well corresponds to the experimental data and give an estimation of the duration of the time unit ($0.004 \sim 10$ days).

In exponential phase (Fig. 12) the lysogenic state (E_x) is practically constant with only a loss of 155 bacteria after 200 days.

To estimate the robustness of these dynamical behaviours with respect to the values of the k_1 and k_4 parameters, we show that in exponential phase k_4 has no incidence and k_1 can vary between 10 and 5000 (Fig. 13), without changing the agreement between predicted and observed dynamical behaviours.

The stability study in the neighbourhood of the stationary states is done by calculating the eigenvalues of the Jacobian matrix of the differential system above: 0 is always an eigenvalue that causes an asymptotic instability (but the system is trajectorially Lyapunov-stable) and, in stationary phase, among all other eigenvalues, two are complex, with negative real part, whereas one is real negative; in exponential phase, two are real negative and one is real positive.

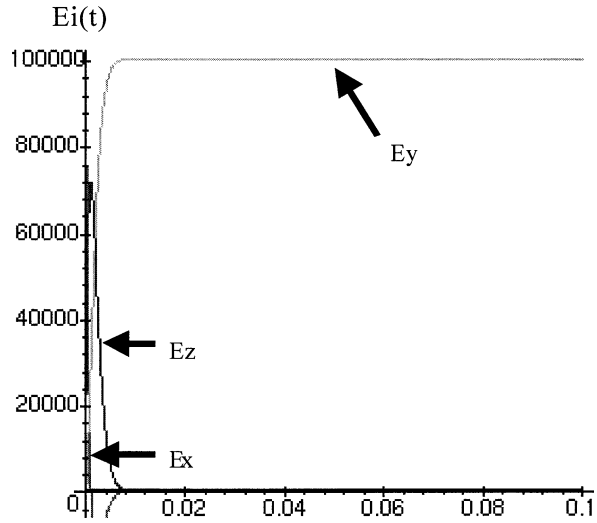


Fig. 11. Simulation of the behaviour of 100 000 Mu lysogenes when bacteria are in stationary phase.

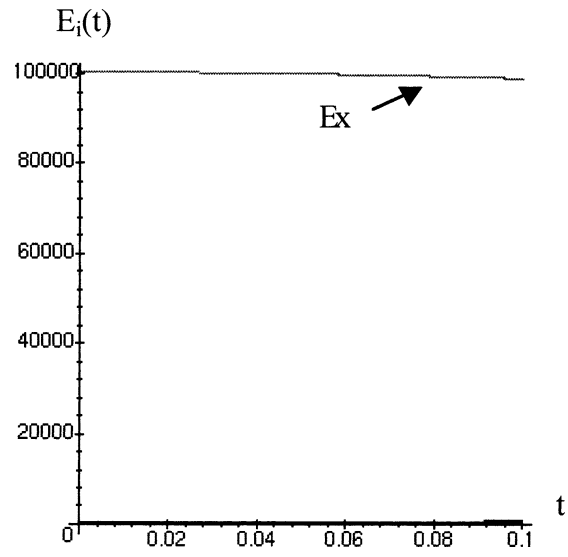


Fig. 12. Simulation of the behaviour of 100 000 Mu lysogenes when bacteria are in exponential phase.

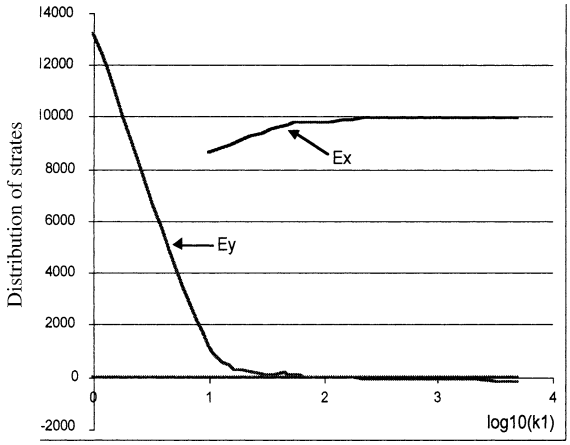


Fig. 13. Evolution of the asymptotic values of E_x and E_y in the exponential phase with respect to the k_1 values.

5.4. Discussion

The Boolean model has shown the necessity to suppress the self-inhibition by Ner in order to get a bi-stability. The continuous version explains in the lytic phase the reaching a plateau behaviour for the 01 state after 10 days and, in the lysogenic phase, the long transient of the 10 state (1% of decrease after 3 months). In order to get a more predictable model, we need measures of the proportion lytic/lysogenic bacteria on a wider interval of growth rates (the local growth rate $R(t)$ coming from a logistic or from a Monod saturation model [68]). Then the system would become non-linear with ISF and IHF being proportionally respectively to $R(t)$ and $\mathrm{d}R(t)/\mathrm{d}t$.

6. Open problems

A frequent criticism made about the Boolean models is their unrealistic number of levels (two); it is possible to easily relax this constraint by considering either a multi-threshold network (for which most of the results of Section 4 are still available), or a non-linear automaton, like the following:

$$x_i(t+1) = \sup\Bigg(-1, \inf\Bigg(1, x_i(t) - sg\Bigg(\sum_j m_{ij}\big(1+x_j(t)\big)\big(1-x_j(t)\big)\Bigg) + sg\Bigg(\sum_j m_{ij}\big(1-x_j(t)\big)\Bigg) + sg\Bigg(\sum_j m_{ij}\big(1+x_j(t)\big)\Bigg)\Bigg)\Bigg)$$

where $sg(y) = 1$, if $y > 0$, $sg(y) = 0$, if $y = 0$ and $sg(y) = -1$, if $y < 0$, or, like in Section 5, a differentiable system. We can also merge the Boolean and the differentiable approach in a hybrid system [69, 70].

Another perspective exists for the interaction matrices introduced above, i.e. the ability to calculate the barycentre between two matrices by using classical (spectral or L_2) distances between matrices, then to build phylogenic trees among a set of species avoiding the complex problems coming from the non-unicity of L_1 (Hamming or Manhattan) barycentres met in the sequence based phylogenic trees. The interaction-based phylogenic trees could reflect more the genomic function than the genomic anatomy and hence could explain more deeply the evolution trends, e.g., by distinguishing the evolution of domestic genetic regulatory networks (involved in rearrangements [12,13]) and that of high-level genetic regulatory networks (corresponding for example to brain signalling or hormonal regulatory systems), the complexity of their interaction matrices being probably of the same order of magnitude than the complexity of the metabolic interaction matrices corresponding to their expressed proteins (carriers, receptors or enzymes) [71,72].

A last important open problem concerns the relationship between the number F of fixed configurations and the number S of interaction loops of the interaction matrix $\mathbf{M}$: the problem is in fact to find the best upper bound for F given an interaction matrix $\mathbf{M}$. This question is the discrete translation of the famous 16th Hilbert's problem of determining an efficient upper bound for the number of limit cycles of a polynomial differential system. Let us summarize the role of the architecture of positive and negative (with an odd number of inhibitions) loops of $\mathbf{M}$ on the occurrence of multiple stationary behaviours as obtained in [8–15]: if the number of genes and the number of interactions are the same, there is only one isolated loop

in **M** and either this loop is negative and the lowest bound (0) for F is reached, or this loop is positive and the upper bound (2^1) for F is reached. If the numbers of genes and interactions are respectively n and $n+1$, there are two interaction loops with the following structure; if both loops are negative, $F=0$; if there are a positive loop and a negative loop disjoint, $F=0$; if there is a positive loop intersecting a negative loop, $F=1$; if there are two disjoint positive loops, $F=2^2$. If, more generally, the number S of loops is m, then: if all loops are negative, $F=0$; if all loops are positive, then: $2 \leqslant F \leqslant 2^m$, and if and only if all loops are positive and disjoint, $F=2^m$. An interesting open problem is now to make exhaustive the determination of F and S and, in particular, to find the circumstances (related to the loops structure) in which we can relate the number of intersecting and isolated loops to F. The approach for solving this open problem could consist first in finding coherent relationships between analogous properties discovered for continuous versions of the regulatory networks and for general Boolean networks.

7. Conclusion

A geneticist could exploit the results given in the above paper as follows: we have shown in Section 4 that it would be possible to characterize the minimal interaction matrices having certain state vectors as fixed configurations. The determination of these matrices is not unique, but permits to focus on certain important equivalence classes, in which the expected matrix has to belong. This considerably restricts the choice of the possible interaction matrices compatible with observed fixed configurations, when it is impossible to directly get from experiments all interaction coefficients, but also when it is only possible to observe the phenomenology of fixed or cyclic configurations. This corresponds in genetics to the phenotypic observation of stationary expression behaviours without experimental measure of the inhibitory and activatory coefficients of promoters and repressors. The possibility to obtain (even in an equivalence class) a sketch of the interaction matrix permits to construct (by randomising the unknown coefficients of **M**) more complicated interaction matrices, then to test if they still have the observed states as fixed configurations, finally keep or reject these tested matrices and propose further experimental strategies, using bio-arrays for refining the knowledge about the genetic network interaction structure.

Acknowledgements

We have done this work thanks to the support of the National Network for Technology Research RNTS 'Technologies for Health' from the French Ministry of Research. We are also indebted to A. Toussaint and C. Ranquet for their helpful discussions concerning the bacteriophage Mu. We thank also B. Hess and A. Winfree (in memoriam) for introducing us to many aspects of the dynamics of life.

References

[1] J. Demongeot, J.-P. Françoise, M. Richard, F. Senegas, T.P. Baum, A differential geometry approach for biomedical image processing, C. R. Biologies 325 (2002) 367–374.
[2] J. Mattes, M. Richard, J. Demongeot, Tree representation for image matching and object recognition, Lect. Notes Comput. Sci. 1568 (1999) 298–309.
[3] L. Mendoza, E.R. Alvarez-Buylla, Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis, J. Theoret. Biol. 193 (1998) 307–319.
[4] F. Thuderoz, DEA Report, UJF, Grenoble, France, 2000.
[5] J. Demongeot, A stochastic model for the cellular metabolism, in: J.R. Barra, et al. (Eds.), Recent Developments in Statistics, North-Holland, Amsterdam, 1977, pp. 655–662.
[6] P.J. Goss, J. Peccoud, Quantitative modeling of stochastic systems in molecular biology using stochastic Petri nets, Proc. Natl Acad. Sci. USA 95 (1998) 6750–6755.
[7] P.J. Goss, J. Peccoud, Analysis of the stabilizing effect of ROM on the genetic network controlling ColE1 plasmid replication, in: R.B. Altam, et al. (Eds.), Pacific Symposium on Biocomputing'99, World Scientific, Singapore, 1999, pp. 65–76.
[8] O. Cinquin, J. Demongeot, Positive and negative feedback: striking a balance between necessary antagonists, J. Theoret. Biol. 216 (2002) 229–241.
[9] O. Cinquin, J. Demongeot, Positive and negative feedback: mending the ways of sloppy systems, C. R. Biologies 325 (2002) 1085–1095.
[10] J. Demongeot, F. Berger, T.-P. Baum, F. Thuderoz, O. Cohen, Bio-array images processing and genetic networks modelling, in: M. Unser, Z.P. Liang (Eds.), ISBI 2002, IEEE EMB, IEEE Proceedings, Piscataway, 2002, pp. 50–54.
[11] J. Demongeot, M. Kaufman, R. Thomas, Interaction matrices, regulation circuits and memory, C. R. Acad. Sci. Paris, Ser. III 323 (2000) 69–80.

[12] J. Demongeot, J. Aracena, S. Ben Lamine, M.A. Mermet, O. Cohen, Hot spots in chromosomal breakage: from description to etiology, in: D. Sankoff, J.-H. Nadeau (Eds.), Comparative Genomics, Kluwer, Amsterdam, 2000, pp. 71–85.
[13] J. Demongeot, J. Aracena, S. Ben Lamine, S. Meignen, A. Tonnelier, R. Thomas, Dynamical systems and biological regulations, in: E. Goles, S. Martinez (Eds.), Complex Systems, Kluwer, Amsterdam, 2000, pp. 107–151.
[14] J. Aracena, J. Demongeot, E. Goles, Fixed points and maximal independent sets on AND-OR networks, Discr. Appl. Math. (in press).
[15] J. Aracena, S. Ben Lamine, M.A. Mermet, O. Cohen, J. Demongeot, Mathematical modelling in genetic networks: relationships between the genetic expression and both chromosomic breakage and positive circuits, in: N. Bourbakis (Ed.), BIBE 2000, IEEE, Piscataway, 2000, pp. 141–149.
[16] R. Thomas, On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations, Springer Ser. Synerget. 9 (1980) 1–23.
[17] M. Delbrück, Discussion, Unités biologiques douées de continuité génétique, Colloques internationaux CNRS 8 (1949) 33–35.
[18] R. Thomas, D. Thieffry, M. Kaufman, Dynamical behavior of biological regulatory networks. I. Biological role and logical analysis of feedback loops, Bull. Math. Biol. 57 (1995) 328–339.
[19] E.H. Snoussi, R. Thomas, Logical identification of all steady states: the concept of feedback loop characteristic states, Bull. Math. Biol. 55 (1993) 973–991.
[20] D. Thieffry, M. Colet, R. Thomas, Formalization of regulatory networks: a logical method and its automatization, Math. Modelling Sci. Comput. 2 (1993) 144–151.
[21] R. Thomas, R. D'Ari, Biological Feedback, CRC Press, Boca Raton, 1990.
[22] S. Kauffman, The Origins of Order, Oxford University Press, Oxford, England, 1993.
[23] C. Berge, Graphes et Hypergraphes, Dunod, Paris, 1974.
[24] E. Goles, S. Martinez, Neural and Automata Networks, in: Maths. and Appl. Series, Vol. 58, Kluwer, Amsterdam, 1991.
[25] F. Plouraboué, H. Atlan, G. Weisbuch, J.-P. Nadal, A network model of the coupling of ion channels with secondary messenger in cell signaling, Network Computation in Neural Networks Systems 3 (1992) 393–406.
[26] B. Bollobas, Random Graphs, Academic Press, London, 1985.
[27] J. Aracena, Modèles mathématiques discrets associés à des systèmes biologiques. Application aux réseaux de régulation génétique, PhD Thesis, U. Chile & UJF, Santiago & Grenoble, 2001.
[28] M. Leptin, Gastrulation in Drosophila: the logic and the cellular mechanisms, EMBO J. 18 (1999) 3187–3192.
[29] N. Rashevsky, Mathematical Biophysics, Cambridge University Press, London, 1948.
[30] A. Turing, The mathematical basis of morphogenesis, Phil. Trans. Roy. Soc. B 237 (1952) 37–47.
[31] A.N. Kolmogorov, I. Petrowski, N. Piskounov, Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique, Mosc. Univ. Bull. Math. 1 (1937) 1–25.
[32] M. Thellier, L. Le Sceller, V. Norris, M.-C. Verdus, C. Ripoll, Long-distance transport, storage and recall of morphogenetic information in plants. The existence of a sort of primitive plant 'memory', C. R. Acad. Sci. Paris, Ser. III 323 (2000) 81–91.
[33] J. Demongeot, M. Thellier, R. Thomas, A mathematical model for storage and recall functions in plants, C. R. Acad. Sci., Ser. III 323 (2000) 93–97.
[34] J. Demongeot, M. Laurent, Sigmoidicity in allosteric models, Math. Biosci. 67 (1983) 1–17.
[35] R. Thomas, La logique des systèmes vivants, Bull. Cl. Sci. Acad. R. Belg. 74 (1988) 432–442.
[36] O. Cinquin, J. Demongeot, Inhibitory n-switch dynamics and applications (submitted).
[37] O. Cohen, M.A. Mermet, J. Demongeot, HC Forum®: a web site based on an international human cytogenetic data base, Nucleic Acids Res. 29 (2001) 305–307.
[38] O. Cohen, M.A. Mermet, J. Demongeot, HC Forum: toward a tele-expertise plat-form in medical genetics, Lect. Notes Med. Inf. 13 (2002) 97–104.
[39] O. Cohen, C. Cans, M. Cuillel, J.-L. Gilardi, H. Roth, M.-A. Mermet, P. Jalbert, J. Demongeot, Cartographic study: breakpoints in 1574 families carrying human reciprocal translocations, Hum. Genet. 97 (1996) 659–667.
[40] O. Cohen, C. Cans, M.-A. Mermet, J. Demongeot, P. Jalbert, Viability thresholds for partial trisomies and monosomies. A study of 1159 viable unbalanced reciprocal translocations, Hum. Genet. 93 (1994) 188–194.
[41] A. Toussaint, M.-J. Gama, J. Laachouch, G. Maenhaut-Michel, Regulation of bacteriophage Mu transposition, Genetica 93 (1994) 27–39.
[42] H. McAdams, L. Shapiro, Circuit simulation of Genetic Networks, Science 269 (1995) 650–656.
[43] A.L. Taylor, Bacteriophage-induced mutation in *E. coli*, Proc. Natl Acad. Sci. USA 50 (1963) 1043–1051.
[44] A.I. Bukhari, D. Zipser, Random insertion of Mu-1 DNA within a single gene, Nat. New Biol. 236 (1972) 240–243.
[45] B. McClintock, Controlling elements and the gene, Cold Spring Harbor Symp. Quant. Biol. 21 (1956) 197–216.
[46] N. Symonds, A. Toussaint, P. van de Putte, M.M. Howe, Phage Mu, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 1987.
[47] C. Wijffelman, M. Gassler, W.F. Stevens, P. Van de Putte, On the control of transcription of bacteriophage Mu, Mol. Gen. Genet. 131 (1974) 85–96.
[48] K. Mizuuchi, In vivo transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction, Cell 35 (1983) 785–794.
[49] C.F. Marrs, M.M. Howe, *Ava*II and *Bgl* I restriction maps of bacteriophage Mu, Virology 126 (1983) 563–575.
[50] H.M. Krause, M.R. Rothwell, N.P. Higgins, The early promoter of bacteriophage Mu: definition of the site of transcript initiation, Nucleic Acids Res. 11 (1983) 5483–5495.
[51] M. Giusti, G. Di Lallo, P. Ghelardini, L. Paolozzi, The bacteriophage Mu Ner gene: a positive regulator of the C operon required for normal levels of late transcription, Virology 179 (1990) 694–700.
[52] K. Mathee, M.M. Howe, Identification of a positive regulator of the Mu middle operon, J. Bacteriol. 172 (1990) 6641–6650.

[53] D.Y. Kwoh, D. Zipser, Specific binding of Mu repressor to DNA, Nature 277 (1979) 489–491.
[54] N. Gossen, P. van de Putte, Role of Ner protein in bacteriophage Mu transposition, J. Bacteriol. 167 (1986) 503–507.
[55] R. Craigie, M. Mizuuchi, K. Mizuuchi, Site specific recognition of the bacteriophage Mu ends by the MuA protein, Cell 39 (1984) 387–394.
[56] M. Mizuuchi, R.A. Weisberg, K. Mizuuchi, DNA sequence of the control region of phage D108: the N-terminal amino acid sequences of repressor and transposase are similar both in phage D108 and its relative, phage Mu, Nucleic Acids Res. 14 (1986) 3813–3825.
[57] M.L. Pato, C. Reich, Instability of transposase activity: evidence from bacteriophage Mu DNA replication, Cell 29 (1982) 219–225.
[58] R. Alazard, M. Bétermier, M. Chandler, *E. coli* IHF stabilises bacteriophage Mu repressor interactions with operator DNA in vitro, Mol. Microbiol. 6 (1992) 1707–1714.
[59] M.-J. Gama, A. Toussaint, N.P. Higgins, Stabilisation of bacteriophage Mu repressor-operator complexes by the *E. coli* IHF protein, Mol. Microbiol. 6 (1992) 1715–1722.
[60] G. Kukolj, M.S. Du Bow, IHF activates the Ner-repressed early promoter of transposable Mu-like phage D108, J. Biol. Chem. 267 (1992) 17827–17835.
[61] M.D. Ditto, D. Roberts, R.A. Weisberg, Growth phase variation of Integration Host Factor level in *E. coli*, J. Bacteriol. 176 (1994) 3738–3748.
[62] M.G. Surette, S.J. Buch, G. Chaconas, Transposomes: stable protein–DNA complexes involved in the in vitro transposition of bacteriophage Mu DNA, Cell 49 (1987) 253–262.
[63] M. Bétermier, I. Poquet, R. Alazard, M. Chandler, Involvement of *E. coli* FIS protein in maintenance of bacteriophage Mu lysogeny by the repressor, J. Bacteriol. 175 (1993) 3798–3811.
[64] M. Bétermier, C. Lefrère, C. Koch, R. Alazard, M. Chandler, The *E. coli* protein Fis: specific binding to the ends of phage Mu DNA and modulation of phage growth, Mol. Microbiol. 3 (1989) 459–468.
[65] C.A. Ball, R. Osuna, K.C. Ferguson, R. Johnson, Dramatic changes in Fis level upon nutrient upshift in *E. coli*, J. Bacteriol. 174 (1992) 8043–8056.
[66] J.F. Thomson, L. Moitoso de Vargas, C. Koch, R. Kahmann, A. Landy, Cellular factors couple recombination with growth phase, Cell 50 (1987) 901–908.
[67] R. Kruklitis, H. Nakai, Participation of bacteriophage Mu A protein and host factors in initiation of Mu DNA synthesis in vitro, J. Biol. Chem. 269 (1994) 16469–16477.
[68] J. Monod, Recherches sur la croissance des cultures bactériennes, Actualités scientifiques et industrielles, Hermann, Paris, 1942.
[69] J. Demongeot, J. Aracena, F. Thuderoz, T.P. Baum, O. Cohen, Genetic regulation networks: circuits, regulons and attractors, C. R. Biologies (in press).
[70] (a) M. Thellier, J. Demongeot, J. Guespin, C. Ripoll, V. Norris, R. Thomas, Storage and recall of environmental signals in a plant: modelling by use of a logical (discrete) formulation (submitted); (b) J. Demongeot, M. Thellier, J. Guespin, C. Ripoll, V. Norris, R. Thomas, Storage and recall of environmental signals in a plant: modelling by use of a differential (discrete) formulation (submitted).
[71] D. Fell, A. Wagner, The small world of metabolism, Nat. Biotechnol. 18 (2000) 1121–1122.
[72] N. Guelzim, S. Bottani, P. Bourgine, F. Képès, Topological and causal structure of the yeast transcriptional regulatory network, Nat. Genet. 31 (2002) 60–63.

RESUME

V(D)J recombination constitutes a somatic site specific DNA recombination, which originates lymphocyte antigen receptor diversity in jawed vertebrates. Concerning the T cell receptor α chains, V and J genes are used from inside the TRA locus toward distal genes during the successive rearrangements, with no allelic exclusion at the genomic level. Experimental quantifications of particular V-J associations were performed in mouse, giving the tendencies of thymic and peripheral combinatorial repertoires. A stochastic numerical model, based on successive opening windows progressing over the V and J regions during the rearrangement rounds, revealed new insights in the understanding of the dynamical rules governing V-J rearrangements and provided a simulated combinatorial repertoire with the entire V-J association frequencies. In the transition to human, thymic quantifications of certain V-J associations were performed, providing a first experimental wide-ranging sampling of the human TRA combinatorial repertoire. The modeling modeling step offered a clear understanding of the dynamical building of the human α repertoire and proposed predictions on repertoire combinatorial diversity richness. Finally, the precise progression of gene accessibility to rearrangements, according to non-constant opening speeds, together with a synchronized opening of the J regions between both alleles, were sufficient to fully explain both the experimental V-J frequencies currently available for the two species as well as the interallelic J usage.

La recombinaison V(D)J constitue une recombinaison somatique et site-spécifique de l'ADN à l'origine de la diversité des récepteurs antigéniques des lymphocytes T chez les vertébrés mandibulés. Concernant la chaîne α des récepteurs T, les gènes V et J sont utilisés depuis l'intérieur du locus TRA vers les gènes distaux durant des réarrangements successifs et ce sans exclusion allélique. La quantification expérimentale de certaines associations V-J chez la souris a permis de définir les tendances des répertoires combinatoires thymiques et périphériques. Un modèle numérique stochastique, basé sur des fenêtrages d'ouverture successives progressant sur les régions V et J durant les cycles de réarrangements, a permis une meilleure compréhension des règles dynamiques gouvernant les réarrangements V-J et a apporté la connaissance d'un répertoire combinatoire simulé renseignant les fréquences de toutes les associations V-J. Lors de la transition à l'homme, la quantification des associations V-J a été réalisée au niveau du thymus, constituant un premier échantillonnage à large échelle du répertoire combinatoire TRA humain. L'étape de modélisation a offert une compréhension claire de la construction dynamique du répertoire α humain et a permis de proposer des prédictions sur la diversité du répertoire combinatoire. Finalement, la progression de l'accessibilité des gènes aux réarrangements selon des vitesses d'ouverture non-constantes associée à une ouverture synchronisée des régions J entre les deux allèles se sont révélées suffisantes pour expliquer les fréquences V-J expérimentales présentement disponibles pour les deux espèces ainsi que l'utilisation interallélique des gènes J.